Coding on countably infinite alphabets

Stéphane Boucheron, Aurélien Garivier and Elisabeth Gassiat

Abstract

This paper describes universal lossless coding strategies for compressing sources on countably infinite alphabets. Classes of memoryless sources defined by an envelope condition on the marginal distribution provide benchmarks for coding techniques originating from the theory of universal coding over finite alphabets. We prove general upper-bounds on minimax regret and lower-bounds on minimax redundancy for such source classes. The general upper bounds emphasize the role of the Normalized Maximum Likelihood codes with respect to minimax regret in the infinite alphabet context. Lower bounds are derived by tailoring sharp bounds on the redundancy of Krichevsky-Trofimov coders for sources over finite alphabets. Up to logarithmic (resp. constant) factors the bounds are matching for source classes defined by algebraically declining (resp. exponentially vanishing) envelopes. Effective and (almost) adaptive coding techniques are described for the collection of source classes defined by algebraically vanishing envelopes. Those results extend our knowledge concerning universal coding to contexts where the key tools from parametric inference are known to fail.

keywords: NML; countable alphabets; redundancy; adaptive compression; minimax;

I. Introduction

This paper is concerned with the problem of universal coding on a countably infinite alphabet $\mathcal{X}$ (say the set of positive integers $\mathbb{N}_+$ or the set of integers $\mathbb{N}$) as described for example by Orlitsky and Santhanam (2004). Throughout this paper, a source on the countable alphabet $\mathcal{X}$ is a probability distribution on the set $\mathcal{X}^{\mathbb{N}}$ of infinite sequences of symbols from $\mathcal{X}$ (this set is endowed with the σ-algebra generated by sets of the form $\prod_{i=1}^{n}\{x_i\}\times\mathcal{X}^{\mathbb{N}}$ where all $x_i \in \mathcal{X}$ and $n \in \mathbb{N}$). The symbol Λ will be used to denote various classes of sources on the countably infinite alphabet $\mathcal{X}$. The sequence of symbols emitted by a source is denoted by the $\mathcal{X}^{\mathbb{N}}$-valued random variable $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$. If P denotes the distribution of $\mathbf{X}$, P^n denotes the distribution of $X_{1:n} = X_1, ..., X_n$, and we let $\Lambda^n = \{P^n : P \in \Lambda\}$. For any countable set $\mathcal{X}$, let $\mathfrak{M}_1(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$.

From Shannon noiseless coding Theorem (see Cover and Thomas, 1991), the binary entropy of P^n, $H(X_{1:n}) = \mathbb{E}_{P^n}\left[-\log P(X_{1:n})\right]$ provides a tight lower bound on the expected number of binary symbols needed to encode outcomes of P^n. Throughout the paper, logarithms are in base 2. In the following, we shall only consider finite entropy sources on countable alphabets, and we implicitly assume that $H(X_{1:n}) < \infty$. The *expected redundancy* of any distribution $Q^n \in \mathfrak{M}_1(\mathcal{X}^n)$, defined as the difference between the expected code length $\mathbb{E}_P\left[-\log Q^n(X_{1:n})\right]$ and $H(X_{1:n})$, is equal to the Kullback-Leibler divergence (or relative entropy) $D(P^n, Q^n) = \sum_{\mathbf{x}\in\mathcal{X}^n} P^n\{\mathbf{x}\}\log\frac{P^n(\mathbf{x})}{Q^n(\mathbf{x})} = \mathbb{E}_{P^n}\left[\log\frac{P^n(X_{1:n})}{Q^n(X_{1:n})}\right]$.

Universal coding attempts to develop sequences of coding probabilities $(Q^n)_n$ so as to minimize expected redundancy over a whole class of sources. Technically speaking, several distinct notions of universality have been

considered in the literature. A positive, real-valued function $\rho(n)$ is said to be a strong (respectively weak) *universal redundancy rate* for a class of sources Λ if there exists a sequence of coding probabilities $(Q_n)_n$ such that for all n, $R^+(Q^n, \Lambda^n) = \sup_{P\in\Lambda} D(P^n, Q^n) \leq \rho(n)$ (respectively for all $P \in \Lambda$, there exists a constant $C(P)$ such that for all n, $D(P^n, Q^n) \leq C(P)\rho(n)$). A redundancy rate $\rho(n)$ is said to be non-trivial if $\lim_n \frac{1}{n}\rho(n) = 0$. Finally a class Λ of sources will be said to be *feebly universal* if there exists a single sequence of coding probabilities $(Q^n)_n$ such that $\sup_{P\in\Lambda} \lim_n \frac{1}{n} D(P^n, Q^n) = 0$ (Note that this notion of feeble universality is usually called weak universality, (see Kieffer, 1978, Györfi et al., 1994), we deviate from the tradition, in order to avoid confusion with the notion of weak universal redundancy rate).

The *maximal redundancy* of Q^n with respect to Λ is defined by:

$$R^+(Q^n, \Lambda^n) = \sup_{P\in\Lambda} D(P^n, Q^n)\,.$$

The infimum of $R^+(Q^n, \Lambda^n)$ is called the *minimax redundancy* with respect to Λ:

$$R^+(\Lambda^n) = \inf_{Q^n\in\mathfrak{M}_1(\mathcal{X}^n)} R^+(Q^n, \Lambda^n).$$

It is the smallest strong universal redundancy rate for Λ. When finite, it is often called the information radius of Λ^n.

As far as finite alphabets are concerned, it is well-known that the class of stationary ergodic sources is feebly universal. This is witnessed by the performance of Lempel-Ziv codes (see Cover and Thomas, 1991). It is also known that the class of stationary ergodic sources over a finite alphabet does not admit any non-trivial weak universal redundancy rate (Shields, 1993). On the other hand, fairly large classes of sources admitting strong universal redundancy rates and non-trivial weak universal redundancy rates have been exhibited (see Barron et al., 1998, Catoni, 2004, and references therein). In this paper, we will mostly focus on strong universal redundancy rates for classes of sources over infinite alphabets. Note that in the latter setting, even feeble universality should not be taken for granted: the class of memoryless processes on $\mathbb{N}_+$ is not feebly universal.

Kieffer (1978) characterized feebly universal classes, and the argument was simplified by Györfi et al. (1994), Gyorfi et al. (1993). Recall that the entropy rate $H(P)$ of a stationary source is defined as $\lim_n H(P^n)/n$. This result may be phrased in the following way.

Proposition 1: A class Λ of stationary sources over a countable alphabet $\mathcal{X}$ is feebly universal if and only if there exists a probability distribution $Q \in \mathfrak{M}_1(\mathcal{X})$ such that for every $P \in \Lambda$ with finite entropy rate, Q satisfies $\mathbb{E}_P \log \frac{1}{Q(X_1)} < \infty$ or equivalently $D(P^1, Q) < \infty$.

Assume that Λ is parameterized by Θ and that Θ can be equipped with (prior) probability distributions W in such a way that $\theta \mapsto P^n_\theta\{A\}$ is a random variable (a measurable mapping) for every $A \subseteq \mathcal{X}^n$. A convenient way to derive lower bounds on $R^+(\Lambda^n)$ consists in using the relation $\mathbb{E}_W[D(P^n_\theta, Q^n)] \leq R^+(Q^n, \Lambda^n)$.

The sharpest lower bound is obtained by optimizing the prior probability distributions, it is called the maximin bound

$$\sup_{W\in\mathfrak{M}_1(\Theta)} \inf_{Q^n\in\mathfrak{M}_1(\mathcal{X}^n)} \mathbb{E}_W[D(P^n_\theta, Q^n)].$$

It has been proved in a series of papers (Gallager, 1968, Davisson, 1973, Haussler, 1997) (and could also have been derived from a general minimax theorem by Sion, 1958) that such a lower bound is tight.

Theorem 1: Let Λ denote a class of sources over some finite or countably infinite alphabet. For each n, the minimax redundancy over Λ coincides with

$$R^+(\Lambda^n) = \sup_{\Theta, W \in \mathfrak{M}_1(\Theta)} \inf_{Q^n \in \mathfrak{M}_1(\mathcal{X}^n)} \mathbb{E}_W \left[D(P^n_\theta, Q^n) \right],$$

where Θ runs over all parameterizations of countable subsets of Λ.

If the set $\Lambda^n = \{P^n \; : \; P \in \Lambda\}$ is not pre-compact with respect to the topology of weak convergence, then both sides are infinite. A thorough account of topological issues on sets of probability measures can be found in (Dudley, 2002). For the purpose of this paper, it is enough to recall that: first, a subset of a metric space is pre-compact if for any $\epsilon > 0$, it can be covered by a finite number of open balls with radius at most $\epsilon > 0$; second, a sequence $(Q_n)_n$ of probability distributions converges with respect to the topology of weak convergence toward the probability distribution Q if and only if for any bounded continuous function h over the support of Q_n's, $\mathbb{E}_{Q_n} h \rightarrow \mathbb{E}_Q h$. This topology can be metrized using the Lévy-Prokhorov distance.

Otherwise the maximin and minimax average redundancies are finite and coincide; moreover, the minimax redundancy is achieved by the mixture coding distribution $Q^n(.) = \int_\Theta P^n_\theta(.) W(\mathrm{d}\theta)$ where W is the least favorable prior.

Another approach to universal coding considers *individual sequences* (see Feder et al., 1992, Cesa-Bianchi and Lugosi, 2006, and references therein). Let the *regret* of a coding distribution Q^n on string $\mathbf{x} \in \mathbb{N}^n_+$ with respect to Λ be $\sup_{P^n \in \Lambda} \log P^n(\mathbf{x})/Q^n(\mathbf{x})$. Taking the maximum with respect to $x \in \mathbb{N}^n_+$, and then optimizing over the choice of Q^n, we get the *minimax regret*:

$$R^*(\Lambda^n) = \inf_{Q^n \in \mathfrak{M}_1(\mathcal{X}^n)} \max_{x \in \mathbb{N}^n_+} \sup_{P \in \Lambda} \log \frac{P^n(x)}{Q^n(x)} .$$

In order to provide proper insight, let us recall the precise asymptotic bounds on minimax redundancy and regret for memoryless sources over finite alphabets (see Clarke and Barron, 1990; 1994, Barron et al., 1998, Xie and Barron, 1997; 2000, Orlitsky and Santhanam, 2004, Catoni, 2004, Szpankowski, 1998, Drmota and Szpankowski, 2004, and references therein).

Theorem 2: Let $\mathcal{X}$ be an alphabet of m symbols, and Λ denote the class of memoryless processes on $\mathcal{X}$ then

$$\lim_n \left\{ R^+(\Lambda^n) - \frac{m-1}{2} \log \frac{n}{2\pi \mathrm{e}} \right\} = \log \left(\frac{\Gamma(1/2)^m}{\Gamma(m/2)} \right)$$
$$\lim_n \left\{ R^*(\Lambda^n) - \frac{m-1}{2} \log \frac{n}{2\pi} \right\} = \log \left(\frac{\Gamma(1/2)^m}{\Gamma(m/2)} \right) .$$

For all $n \geq 2$:

$$R^*(\Lambda^n) \leq \frac{m-1}{2} \log n + 2 .$$

The last inequality is checked in the Appendix .

Remark 1: The phenomenon pointed out in Theorem 2 holds not only for the class of memoryless sources over a

finite alphabet but also for classes of sources that are smoothly parameterized by finite dimensional sets (see again Clarke and Barron, 1990; 1994, Barron et al., 1998, Xie and Barron, 1997; 2000, Orlitsky and Santhanam, 2004, Catoni, 2004).

The minimax regret deserves further attention. For a source class Λ, for every $\mathbf{x} \in \mathcal{X}^n$, let the maximum likelihood $\hat{p}(\mathbf{x})$ be defined as $\sup_{P \in \Lambda} P^n(\mathbf{x})$. If $\sum_{\mathbf{x} \in \mathbb{N}_+^n} \hat{p}(\mathbf{x}) < \infty$, the *Normalized Maximum Likelihood* coding probability is well-defined and given by

$$Q^n_{\text{NML}}(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})}{\sum_{x \in \mathbb{N}_+^n} \hat{p}(\mathbf{x})} .$$

Shtarkov (1987) showed that the *Normalized Maximum Likelihood* coding probability achieves the same regret over all strings of length n and that this regret coincides with the *minimax regret*:

$$R^*(\Lambda^n) = \log \sum_{\mathbf{x} \in \mathbb{N}_+^n} \hat{p}(\mathbf{x}).$$

Memoryless sources over finite alphabets are special cases of envelope classes. The latter will be of primary interest.

Definition 1: Let f be a mapping from $\mathbb{N}_+$ to $[0,1]$. The envelope class Λ_f defined by function f is the collection of stationary memoryless sources with first marginal distribution dominated by f:

$$\Lambda_f = \left\{ P \ : \ \ \forall x \in \mathbb{N},\ P^1\{x\} \leq f(x) \ , \text{ and } P \text{ is stationary and memoryless.} \right\} .$$

We will be concerned with the following topics.

1) Understanding general structural properties of minimax redundancy and minimax regret.
2) Characterizing those source classes that have finite minimax regret.
3) Quantitative relations between minimax redundancy or regret and integrability of the envelope function.
4) Developing effective coding techniques for source classes with known non-trivial minimax redundancy rate.
5) Developing adaptive coding schemes for collections of source classes that are too large to enjoy even a weak redundancy rate.

The paper is organized as follows. Section II describes some structural properties of minimax redundancies and regrets for classes of stationary memoryless sources. Those properties include monotonicity and sub-additivity. Proposition 5 characterizes those source classes that admit finite regret. This characterization emphasizes the role of Shtarkov Normalized Maximum Likelihood coding probability. Proposition 6 describes a simple source class for which the minimax regret is infinite, while the minimax redundancy is finite. Finally Proposition 3 asserts that such a contrast is not possible for the so-called envelope classes.

In Section III, Theorems 4 and 5 provide quantitative relations between the summability properties of the envelope function and minimax regrets and redundancies. Those results build on the non-asymptotic bounds on minimax redundancy derived by Xie and Barron (1997).

Section IV focuses on two kinds of envelope classes. This section serves as a benchmark for the two main

results from the preceding section. In Subsection IV-A, lower-bounds on minimax redundancy and upper-bounds on minimax regret for classes defined by envelope function $k \mapsto 1 \wedge Ck^{-\alpha}$ are described. Up to a factor $\log n$ those bounds are matching. In Subsection IV-B, lower-bounds on minimax redundancy and upper-bounds on minimax regret for classes defined by envelope function $k \mapsto 1 \wedge C \exp^{-\alpha k}$ are described. Up to a multiplicative constant, those bounds coincide and grow like $\log^2 n$.

In Sections V and VI, we turn to effective coding techniques geared toward source classes defined by power-law envelopes. In Section V, we elaborate on the ideas embodied in Proposition 4 from Section II, and combine mixture coding and Elias penultimate code (Elias, 1975) to match the upper-bounds on minimax redundancy described in Section IV. One of the messages from Section IV is that the union of envelope classes defined by power laws, does not admit a weak redundancy rate that grows at a rate slower than $n^{1/\beta}$ for any $\beta > 1$. In Section VI, we finally develop an adaptive coding scheme for the union of envelope classes defined by power laws. This adaptive coding scheme combines the censoring coding technique developed in the preceding subsection and an estimation of tail-heaviness. It shows that the union of envelope classes defined by power laws is feebly universal.

II. STRUCTURAL PROPERTIES OF THE MINIMAX REDUNDANCY AND MINIMAX REGRET

Propositions 2,3 and 4 below are sanity-check statements: they state that when minimax redundancies and regrets are finite, as functions of word-length, they are non-decreasing and sub-additive. In order to prove them, we start by the following proposition which emphasizes the role of the NML coder with respect to the minimax regret. At best, it is a comment on Shtarkov's original work (Shtarkov, 1987, Haussler and Opper, 1997).

Proposition 2: Let Λ be a class of stationary memoryless sources over a countably infinite alphabet, the minimax regret with respect to Λ^n, $R^*(\Lambda^n)$ is finite if and only if the normalized maximum likelihood (Shtarkov) coding probability Q^n_{NML} is well-defined and given by

$$Q^n_{\text{NML}}(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})}{\sum_{\mathbf{y} \in \mathcal{X}^n} \hat{p}(\mathbf{y})} \text{ for } \mathbf{x} \in \mathcal{X}^n$$

where $\hat{p}(\mathbf{x}) = \sup_{P \in \Lambda} P^n(\mathbf{x})$.

Note that the definition of Q^n_{NML} does not assume either that the maximum likelihood is achieved on Λ or that it is uniquely defined.

Proof: The fact that if Q^n_{NML} is well-defined, the minimax regret is finite and equal to

$$\log \left(\sum_{\mathbf{y} \in \mathcal{X}^n} \hat{p}(\mathbf{y}) \right)$$

is the fundamental observation of Shtarkov (1987).

On the other hand, if $R^*(\Lambda^n) < \infty$, there exists a probability distribution Q^n on $\mathcal{X}^n$ and a finite number r such that for all $\mathbf{x} \in \mathcal{X}^n$,

$$\hat{p}(\mathbf{x}) \leq r \times Q^n(\mathbf{x}) \,,$$

summing over $\mathbf{x}$ gives

$$\sum_{\mathbf{x}\in\mathcal{X}^n} \hat{p}(\mathbf{x}) \le r < \infty\,.$$

■

Proposition 3: Let Λ denote a class of sources, then the minimax redundancy $R^+(\Lambda^n)$ and the minimax regret $R^*(\Lambda^n)$ are non-decreasing functions of n.

Proof: As far as R^+ is concerned, by Theorem 1, it is enough to check that the maximin (mutual information) lower bound is non-decreasing.

For any prior distribution W on a parameter set Θ (recall that $\{P_\theta \,:\, \theta\in\Theta\}\subseteq\Lambda$, and that the mixture coding probability Q^n is defined by $Q^n(A) = \mathbb{E}_W[P^n_\theta(A)]$)

$$\mathbb{E}_W\left[D(P_\theta^{n+1}, Q^{n+1})\right] = I(\theta; X_{1:n+1}) = I\left(\theta; (X_{1:n}, X_{n+1})\right) \ge I(\theta; X_{1:n}) = \mathbb{E}_W\left[D(P_\theta^n, Q^n)\right].$$

Let us now consider the minimax regret. It is enough to consider the case where $R^*(\Lambda^n)$ is finite. Thus we may rely on Proposition 2. Let n and m be two positive integers. Let ϵ be a small positive real. For any string $\mathbf{x}\in\mathcal{X}^n$, let $P_{\mathbf{x}}\in\Lambda$, be such that $P_{\mathbf{x}}\{\mathbf{x}\} \ge \hat{p}(\mathbf{x})(1-\epsilon)$. Then

$$\begin{aligned}\hat{p}(\mathbf{x}x') &\ge P_{\mathbf{x}}(\mathbf{x}) \times P_{\mathbf{x}}(x' \mid \mathbf{x})\\ &\ge \hat{p}(\mathbf{x})(1-\epsilon) \times P_{\mathbf{x}}(x' \mid \mathbf{x})\,.\end{aligned}$$

Summing over all possible $x'\in\mathcal{X}$ we get

$$\sum_{x'} \hat{p}(\mathbf{x}x') \ge \hat{p}(\mathbf{x})(1-\epsilon)\,.$$

Summing now over all $\mathbf{x}\in\mathcal{X}^n$ and $x'\in\mathcal{X}$,

$$\sum_{\mathbf{x}\in\mathcal{X}^n, x'\in\mathcal{X}} \hat{p}(\mathbf{x}x') \ge \sum_{\mathbf{x}\in\mathcal{X}^n} \hat{p}(\mathbf{x})(1-\epsilon)\,.$$

So that by letting ϵ tend to 0,

$$\sum_{\mathbf{x}\in\mathcal{X}^{n+1}} \hat{p}(\mathbf{x}) \ge \sum_{\mathbf{x}\in\mathcal{X}^n} \hat{p}(\mathbf{x})\,.$$

■

Note that the proposition holds even though Λ is not a collection of memoryless sources. This Proposition can be easily completed when dealing with memoryless sources.

Proposition 4: If Λ is a class of stationary memoryless sources, then the functions $n\mapsto R^+(\Lambda^n)$ and $n\mapsto R^*(\Lambda^n)$ are either infinite or sub-additive.

Proof: Assume that $R^+(\Lambda^n) < \infty$. Here again, given Theorem 1, in order to establish sub-additivity for R^+, it is enough to check the property for the maximin lower bound. Let n, m be two positive integers, and W be any prior on Θ (with $\{P_\theta \,:\, \theta\in\Theta\}\subseteq\Lambda$). As sources from Λ are memoryless, $X_{1:n}$ and $X_{n+1:n+m}$ are independent

conditionally on θ and thus

$$\begin{aligned}
& I\left(X_{n+1:n+m};\theta|X_{1:n}\right) \\
& = H\left(X_{n+1:n+m}|X_{1:n}\right) - H\left(X_{n+1:n+m}|X_{1:n},\theta\right) \\
& = H\left(X_{n+1:n+m}|X_{1:n}\right) - H\left(X_{n+1:n+m}|\theta\right) \\
& \leq H\left(X_{n+1:n+m}\right) - H\left(X_{n+1:n+m}|\theta\right) \\
& = I\left(X_{n+1:n+m};\theta\right) \, .
\end{aligned}$$

Hence, using the fact that under each P_θ, the process $(X_n)_{n\in\mathbb{N}_+}$ is stationary:

$$\begin{aligned}
I\left(X_{1:n+m};\theta\right) &= I\left(X_{1:n};\theta\right) + I\left(X_{n+1:n+m};\theta|X_{1:n}\right) \\
&\leq I\left(X_{1:n};\theta\right) + I\left(X_{n+1:n+m};\theta\right) \\
&= I\left(X_{1:n};\theta\right) + I\left(X_{1:m};\theta\right).
\end{aligned}$$

Let us now check the sub-additivity of the minimax regret. Suppose that $R^*(\Lambda^1)$ is finite. For any $\epsilon > 0$, for $\mathbf{x} \in \mathcal{X}^{n+m}$, let $P \in \Lambda$ be such that $(1-\epsilon)\hat{p}(\mathbf{x}) \leq P^{n+m}(\mathbf{x})$. As for $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{x}' \in \mathcal{X}^m$, $P^{n+m}(\mathbf{x}\mathbf{x}') = P^n(\mathbf{x}) \times P^m(\mathbf{x}')$, we have for any $\epsilon > 0$, and any $\mathbf{x} \in \mathcal{X}^n, \mathbf{x}' \in \mathcal{X}^m$

$$(1-\epsilon)\hat{p}(\mathbf{x}\mathbf{x}') \leq \hat{p}(\mathbf{x}) \times \hat{p}(\mathbf{x}')\,.$$

Hence, letting ϵ tend to 0, and summing over all $\mathbf{x} \in \mathcal{X}^{n+m}$:

$$\begin{aligned}
& R^*\left(\Lambda^{n+m}\right) \\
& = \log \sum_{\mathbf{x}\in\mathcal{X}^n \mathbf{x}'\in\mathcal{X}^m} \hat{p}\left(\mathbf{x}\mathbf{x}'\right) \\
& \leq \log \sum_{\mathbf{x}\in\mathcal{X}^n} \hat{p}(\mathbf{x}) + \log \sum_{\mathbf{x}\in\mathcal{X}^m} \hat{p}\left(\mathbf{x}'\right) \\
& = R^*\left(\Lambda^n\right) + R^*\left(\Lambda^m\right).
\end{aligned}$$

■

Remark 2: Counter-examples witness the fact that subadditivity of redundancies does not hold in full generality. The Fekete Lemma (see Dembo and Zeitouni, 1998) leads to:

Corollary 1: Let Λ denote a class of stationary memoryless sources over a countable alphabet. For both minimax redundancy R^+ and minimax regret R^*,

$$\lim_{n\to\infty} \frac{R^+\left(\Lambda^n\right)}{n} = \inf_{n\in\mathbb{N}_+} \frac{R^+\left(\Lambda^n\right)}{n} \leq R^+\left(\Lambda^1\right) \, ,$$

and

$$\lim_{n\to\infty} \frac{R^*\left(\Lambda^n\right)}{n} = \inf_{n\in\mathbb{N}_+} \frac{R^*\left(\Lambda^n\right)}{n} \leq R^*\left(\Lambda^1\right) \, .$$

Hence, in order to prove that $R^+(\Lambda^n) < \infty$ (respectively $R^*(\Lambda^n) < \infty$), it is enough to check that $R^+\left(\Lambda^1\right) < \infty$ (respectively $R^*\left(\Lambda^1\right) < \infty$).

The following Proposition combines Propositions 2, 3 and 4. It can be rephrased as follows: a class of memoryless sources admits a non-trivial strong minimax regret if and only if Shtarkov NML coding probability is well-defined for $n = 1$.

Proposition 5: Let Λ be a class of stationary memoryless sources over a countably infinite alphabet. Let $\hat{p}$ be defined by $\hat{p}(x) = \sup_{P\in\Lambda} P\{x\}$. The minimax regret with respect to Λ^n is finite if and only if the normalized maximum likelihood (Shtarkov) coding probability is well-defined and :

$$R^*(\Lambda^n) < \infty \Leftrightarrow \sum_{x\in\mathbb{N}_+} \hat{p}(x) < \infty.$$

Proof: The direct part follows from Proposition 2.

For the converse part, if $\sum_{x\in\mathbb{N}_+} \hat{p}(x) = \infty$, then $R^*(\Lambda^1) = \infty$ and from Proposition 3, $R^*(\Lambda^n) = \infty$ for every positive integer n. ∎

When dealing with smoothly parameterized classes of sources over finite alphabets (see Barron et al., 1998, Xie and Barron, 2000) or even with the massive classes defined by renewal sources (Csiszár and Shields, 1996), the minimax regret and minimax redundancy are usually of the same order of magnitude (see Theorem 2 and comments in the Introduction). This can not be taken for granted when dealing with classes of stationary memoryless sources over a countable alphabet.

Proposition 6: Let f be a positive, strictly decreasing function defined on $\mathbb{N}$ such that $f(1) < 1$. For $k \in \mathbb{N}$, let p_k be the probability mass function on $\mathbb{N}$ defined by:

$$p_k(l) = \begin{cases} 1 - f(k) & \text{if } l = 0; \\ f(k) & \text{if } l = k; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Lambda^1 = \{p_1, p_2, \ldots\}$, let Λ be the class of stationary memoryless sources with first marginal Λ^1. The finiteness of the minimax redundancy with respect to Λ^n depends on the limiting behavior of $f(k)\log k$: for every positive integer n:

$$f(k)\log k \to_{k\to\infty} \infty \Leftrightarrow R^+(\Lambda^n) = \infty\,.$$

Remark 3: When $f(k) = \frac{1}{\log k}$, the minimax redundancy $R^+(\Lambda^n)$ is finite for all n. Note, however that this does not warrant the existence of a non-trivial strong universal redundancy rate. However, as $\sum_k f(k) = \infty$, minimax regret is infinite by Proposition 5.

A similar result appears in the discussion of Theorem 3 in (Haussler and Opper, 1997) where classes with finite minimax redundancy and infinite minimax regret are called irregular.

We will be able to refine those observations after the statement of Corollary 2.

Proof: Let us first prove the direct part. Assume that $f(k) \log k \to_{k\to\infty} \infty$. In order to check that $R^+(\Lambda^1) = \infty$, we resort to the mutual information lower bound (Theorem 1) and describe an appropriate collection of Bayesian games.

Let m be a positive integer and let θ be uniformly distributed over $\{1, 2, \ldots, m\}$. Let X be distributed according to p_k conditionally on $\theta = k$. Let Z be the random variable equal to 1 if $X = \theta$ and equal to 0 otherwise. Obviously, $H(\theta|X, Z = 1) = 0$; moreover, as f is assumed to be non-increasing, $P(Z = 0|\theta = k) = 1 - f(k) \leq 1 - f(m)$ and thus:

$$\begin{aligned} H(\theta|X) &= H(Z|X) + H(\theta|Z, X) \\ &\leq 1 + P(Z = 0)H(\theta|X, Z = 0) \\ &\quad +P(Z = 1)H(\theta|X, Z = 1) \\ &\leq 1 + (1 - f(m)) \log m. \end{aligned}$$

Hence,

$$\begin{aligned} R^+(\Lambda^1) &\geq I(\theta, X) \\ &\geq \log m - (1 - f(m)) \log m \\ &= f(m) \log m \end{aligned}$$

which grows to infinity with m, so that as announced $R^+(\Lambda^1) = \infty$.

Let us now prove the converse part. Assume that the sequence $(f(k) \log k)_{k\in\mathbb{N}_+}$ is upper-bounded by some constant C. In order to check that $R^+(\Lambda^n) < \infty$, for all n, by Proposition 4, it is enough to check that $R^+(\Lambda^1_f) < \infty$, and thus, it is enough to exhibit a probability distribution Q over $\mathcal{X} = \mathbb{N}$ such that $\sup_{P\in\Lambda^1} D(P, Q) < \infty$.

Let Q be defined by $Q(k) = A/((1 \vee (k(\log k)^2))$ for $k \geq 2$, $Q(0), Q(1) > 0$ where A is a normalizing constant that ensures that Q is a probability distribution over $\mathcal{X}$.

Then for any $k \geq 3$ (which warrants $k(\log k)^2 > 1$), letting P_k be the probability defined by the probability mass function p_k:

$$
\begin{aligned}
&D(P_k, Q) \\
&= (1 - f(k)) \log \frac{(1 - f(k))}{Q(0)} + f(k) \log \left(\frac{f(k)k(\log k)^2}{A} \right) \\
&\leq -\log Q(0) + C + f(k) \left(2 \log^{(2)}(k) - \log(A) \right) \\
&\leq C + \log \frac{C^2}{A\, Q(0)} \,.
\end{aligned}
$$

This is enough to conclude that

$$
R^+(\Lambda^1) \leq \left(C + \log \frac{C^2}{A\, Q(0)} \right) \vee D(P_1, Q) \vee D(P_2, Q) < \infty \,.
$$

■

Remark 4: Note that the coding probability used in the proof of the converse part of the proposition corresponds to one of the simplest prefix codes for integers proposed by Elias (1975).

The following theorem shows that, as far as envelope classes are concerned (see Definition 1), minimax redundancy and minimax regret are either both finite of both infinite. This is indeed much less precise than the relation stated in Theorem 2 about classes of sources on finite alphabets.

Theorem 3: Let f be a non-negative function from $\mathbb{N}_+$ to $[0, 1]$, let Λ_f be the class of stationary memoryless sources defined by envelope f. Then

$$
R^+ \left(\Lambda_f^n \right) < \infty \Leftrightarrow R^* \left(\Lambda_f^n \right) < \infty \,.
$$

Remark 5: We will refine this result after the statement of Corollary 2.

Recall from Proposition 5 that $R^* \left(\Lambda_f^n \right) < \infty \Leftrightarrow \sum_{k \in \mathbb{N}_+} f(k) < \infty$.

Proof:

In order to check that

$$
\sum_{k \in \mathbb{N}_+} f(k) = \infty \Rightarrow R^+ \left(\Lambda_f^n \right) = \infty \,,
$$

it is enough to check that if $\sum_{k \in \mathbb{N}_+} f(k) = \infty$, the envelope class contains an infinite collection of mutually singular sources.

Let the infinite sequence of integers $(h_i)_{i \in \mathbb{N}}$ be defined recursively by $h_0 = 0$ and

$$
h_{i+1} = \min \left\{ h \ : \ \sum_{k = h_i + 1}^{h} f(k) > 1 \right\} .
$$

The memoryless source P_i is defined by its first marginal P_i^1 which is given by

$$P_i^1(m) = \frac{f(m)}{\sum_{k=h_i+1}^{h_{i+1}} f(k)} \text{ for } m \in \{p_i + 1, ..., p_{i+1}\} .$$

Taking any prior with infinite Shannon entropy over the $\{P_i^1 \ ; \ i \in \mathbb{N}_+\}$ shows that

$$R^+ \left(\{P_i^1 \ ; \ i \in \mathbb{N}_+\}\right) = \infty .$$

■

III. Envelope classes

The next two theorems establish quantitative relations between minimax redundancy and regrets and the shape of the envelope function. Even though the two theorems deal with general envelope functions, the reader might appreciate to have two concrete examples of envelope in mind: exponentially decreasing envelopes of the form $Ce^{-\alpha k}$ for appropriate α and C, and power-laws of the form $Ck^{-\alpha}$ again for appropriate α and C. The former family of envelope classes extends the class of sources over finite (but unknown) alphabets. The first theorem holds for any class of memoryless sources.

Theorem 4: If Λ is a class of memoryless sources, let the tail function $\bar{F}_{\Lambda^1}$ be defined by $\bar{F}_{\Lambda^1}(u) = \sum_{k>u} \hat{p}(k)$, then:

$$R^*(\Lambda^n) \leq \inf_{u : u \leq n} \left[n\bar{F}_{\Lambda^1}(u) \log e + \frac{u-1}{2} \log n \right] + 2 .$$

Choosing a sequence $(u_n)_n$ of positive integers in such a way that $u_n \to \infty$ while $(u_n \log n)/n \to 0$, this theorem allows to complete Proposition 5.

Corollary 2: Let Λ denote a class of memoryless sources, then the following holds:

$$R^*(\Lambda^n) < \infty \Leftrightarrow R^*(\Lambda^n) = o(n) \text{ and } R^+(\Lambda^n) = o(n) .$$

Remark 6: We may now have a second look at Proposition 6 and Theorem 3. In the setting of Proposition 6, this Corollary asserts that if $\sum_k f(k) < \infty$, for the source class defined by f, a non-trivial strong redundancy rate exists.

On the other hand, this corollary complements Theorem 3 by asserting that envelope classes have either non-trivial strong redundancy rates or infinite minimax redundancies.

Remark 7: Again, this statement has to be connected with related propositions from Haussler and Opper (1997). The last paper establishes bounds on minimax redundancy using geometric properties of the source class under

Hellinger metric. For example, Theorem 4 in (Haussler and Opper, 1997) relates minimax redundancy and the metric dimension of the set Λ^n with respect to the Hellinger metric (which coincides with L_2 metric between the square roots of densities) under the implicit assumption that sources lying in small Hellinger balls have finite relative entropy (so that upper bounds in Lemma 7 there are finite). Envelope classes may not satisfy this assumption. Hence, there is no easy way to connect Theorem 4 and results from (Haussler and Opper, 1997).

Proof: (Theorem 4.) Any integer u defines a decomposition of a string $\mathbf{x} \in \mathbb{N}_+^n$ into two non-contiguous substrings: a substring $\mathbf{z}$ made of the m symbols from $\mathbf{x}$ that are larger than u, and one substring $\mathbf{y}$ made of the $n-m$ symbols that are smaller than u.

$$
\begin{aligned}
&\sum_{\mathbf{x}\in\mathbb{N}_+^n} \hat{p}(\mathbf{x}) \\
&\stackrel{(a)}{=} \sum_{m=0}^{n} \binom{n}{m} \sum_{\mathbf{z}\in\{u+1,\dots\}^m} \sum_{\mathbf{y}\in\{1,2,\dots,u\}^{n-m}} \hat{p}\left(\mathbf{z}\mathbf{y}\right) \\
&\stackrel{(b)}{\leq} \sum_{m=0}^{n} \binom{n}{m} \sum_{\mathbf{z}\in\{u+1,\dots\}^m} \prod_{i=1}^{m} \hat{p}\left(\mathbf{z}_i\right) \sum_{\mathbf{y}\in\{1,2,\dots,u\}^{n-m}} \hat{p}\left(\mathbf{y}\right) \\
&\stackrel{(c)}{\leq} \left(\sum_{m=0}^{n} \binom{n}{m} \bar{F}_{\Lambda^1}(u)^m\right) \left(\sum_{\mathbf{y}\in\{1,2,\dots,u\}^{n}} \hat{p}\left(\mathbf{y}\right)\right) \\
&\stackrel{(d)}{\leq} \left(1+\bar{F}_{\Lambda^1}(u)\right)^n 2^{\frac{u-1}{2}\log n+2}.
\end{aligned}
$$

Equation (a) is obtained by reordering the symbols in the strings, Inequalities (b) and (c) follow respectively from Proposition 4 and Proposition 3. Inequality (d) is a direct consequence of the last inequality in Theorem 2.

Hence,

$$
\begin{aligned}
R^*(\Lambda^n) &\leq n\log\left(1+\bar{F}_{\Lambda^1}(u)\right) + \frac{u-1}{2}\log n + 2 \\
&\leq n\bar{F}_{\Lambda^1}(u)\log e + \frac{u-1}{2}\log n + 2
\end{aligned}
$$

■

The next theorem complements the upper-bound on minimax regret for envelope classes (Theorem 4). It describes a general lower bound on minimax redundancy for envelope classes.

Theorem 5: Let f denote a non-increasing, summable envelope function. For any integer p, let $c(p)=\sum_{k=1}^{p} f(2k)$. Let $c(\infty)=\sum_{k\geq 1} f(2k)$. Assume furthermore that $c(\infty)>1$. Let $p\in\mathbb{N}_+$ be such that $c(p)>1$. Let $n\in\mathbb{N}_+$, $\epsilon>0$ and $\lambda\in]0,1[$ be such that $n>\frac{c(p)}{f(2p)}\frac{10}{\epsilon(1-\lambda)}$. Then

$$
R^+(\Lambda_f^n) \geq C(p,n,\lambda,\epsilon)\sum_{i=1}^{p}\left(\frac{1}{2}\log\frac{n\,(1-\lambda)\,\pi f(2i)}{2c(p)e} - \epsilon\right),
$$

where $C(p,n,\lambda,\epsilon)=\frac{1}{1+\frac{c(p)}{\lambda^2 n f(2p)}}\left(1-\frac{4}{\pi}\sqrt{\frac{5c(p)}{(1-\lambda)\epsilon n f(2p)}}\right)$.

Before proceeding to the proof, let us mention the following non-asymptotic bound from Xie and Barron (1997). Let m_n^* denote the Krichevsky-Trofimov distribution over $\{0,1\}^n$. That is, for any $\mathbf{x} \in \{0,1\}^n$, such that $n_1 = \sum_{i=1}^n \mathbf{x}_i$ and $n_0 = n - n_1$

$$m_n^*(\mathbf{x}) = \frac{1}{\pi}\int_{[0,1]} \theta^{n_1-1/2}(1-\theta)^{n_0-1/2}\mathrm{d}\theta.$$

Lemma 1: (Xie and Barron, 1997, Lemma 1) For any $\varepsilon > 0$, there exists a $c(\varepsilon)$ such that for $n > 2c(\varepsilon)$ the following holds uniformly over $\theta \in [c(\varepsilon)/n, 1 - c(\varepsilon)/n]$:

$$\left| D\left(p_\theta^n, m_n^*\right) - \frac{1}{2}\log\frac{n}{2\pi\mathrm{e}} - \log\pi \right| \leqslant \varepsilon.$$

The bound $c(\varepsilon)$ can be chosen as small as $5/\varepsilon$.

Proof: The proof is organized in the following way. A prior probability distribution is first designed in such a way that it is supported by probability distributions that satisfy the envelope condition, have support equal to $\{1,,\ldots,2p\}$, and most importantly enjoys the following property. Letting $\mathbf{N}$ be the random vector from $\mathbb{N}^p$ defined by $\mathbf{N}_i(\mathbf{x}) = |\{j \;:\; \mathbf{x}_j \in \{2i-1, 2i\}\}|$, where $\mathbf{x} \in \mathcal{X}^n$, letting Q^* be the mixture distribution over $\mathcal{X}^n$ defined by the prior, then for any P in the support of the prior, $P^n/Q^* = P^n\{. \mid \mathbf{N}\}/Q^*\{. \mid \mathbf{N}\}$. This property will provide a handy way to bound the capacity of the channel.

Let f, n, p, ϵ and λ be as in the statement of the theorem. Let us first define a prior probability on Λ_f^1. For each integer i between 1 and p, let μ_i be defined as

$$\mu_i = \frac{f(2i)}{c(p)}.$$

This ensures that the sequence $(\mu_i)_{i\le p}$ defines a probability mass function over $\{1,\ldots,p\}$. Let $\boldsymbol{\theta} = (\theta_i)_{1\le i\le p}$ be a collection of independent random variables each distributed according to a Beta distribution with parameters $(1/2, 1/2)$. The prior probability $W = \otimes_{i=1}^p W_i$ for $\boldsymbol{\theta}$ on $[0,1]^p$ has thus density w given by

$$w(\boldsymbol{\theta}) = \frac{1}{\pi^p}\prod_{i=1}^p \left(\theta_i^{-1/2}(1-\theta_i)^{-1/2}\right).$$

The memoryless source parameterized by $\boldsymbol{\theta}$ is defined by the probability mass function $p_{\boldsymbol{\theta}}(2i-1) = \theta_i\,\mu_i$ and $p_{\boldsymbol{\theta}}(2i) = (1-\theta_i)\mu_i$ for $i :\; 1 \le i \le p$ and $p_{\boldsymbol{\theta}}(j) = 0$ for $j > 2i$. Thanks to the condition $c(p) > 1$, this probability mass function satisfies the envelope condition.

For $i \le p$, let the random variable N_i (resp. N_i^0) be defined as the number of occurrences of $\{2i-1, 2i\}$ (resp. $2i-1$) in the sequence $\mathbf{x}$. Let $\mathbf{N}$ (resp. $\mathbf{N}^0$) denote the random vector $N_1, \ldots, N_p$ (resp. $N_1^0, \ldots, N_p^0$). If a sequence $\mathbf{x}$ from $\{1,\ldots,2p\}^n$ contains $n_i = N_i(\mathbf{x})$ symbols from $\{2i-1, 2i\}$ for each $i \in \{1,\ldots,p\}$, and if for each such i, the sequence contains $n_i^0 = N_i^0(\mathbf{x})$ (n_i^1) symbols equal to $2i-1$ (resp. $2i$) then

$$P_{\boldsymbol{\theta}}^n(\mathbf{x}) = \prod_{i=1}^p \left(\mu_i^{n_i}\,\theta_i^{n_i^0}\,(1-\theta_i)^{n_i^1}\right).$$

Note that when the source $\boldsymbol{\theta}$ is picked according to the prior W and the sequence $X_{1:n}$ picked according to $P^n_{\boldsymbol{\theta}}$, the random vector $\mathbf{N}$ is multinomially distributed with parameters n and $(\mu_1, \mu_2, \ldots, \mu_p)$, so the distribution of $\mathbf{N}$ does not depend on the outcome of $\boldsymbol{\theta}$. Moreover, conditionally on $\mathbf{N}$, the conditional probability $P^n_{\boldsymbol{\theta}}\{\cdot \mid \mathbf{N}\}$ is a product distribution:

$$P^n_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{N}) = \prod_{i=1}^{p} \left(\theta_i^{n_i^0} (1-\theta_i)^{n_i^1} \right) .$$

In statistical parlance, the random vectors $\mathbf{N}$ and $\mathbf{N}^0$ form a sufficient statistic for $\boldsymbol{\theta}$.

Let Q^* denote the mixture distribution on $\mathbb{N}_+^n$ induced by W:

$$Q^*(\mathbf{x}) = \mathbb{E}_W \left[P^n_{\boldsymbol{\theta}}(\mathbf{x}) \right] ,$$

and, for each n, let m^*_n denote the Krichevsky-Trofimov mixture over $\{0,1\}^n$, then

$$Q^*(\mathbf{x}) = \prod_{i=1}^{p} \left(\mu_i^{n_i} \, m^*_{n_i}(0^{n_i^0} 1^{n_i^1}) \right) .$$

For a given value of $\mathbf{N}$, the conditional probability $Q^*\{\cdot \mid \mathbf{N}\}$ is also a product distribution:

$$Q^*(\mathbf{x} \mid \mathbf{N}) = \prod_{i=1}^{p} m^*_{N_i}(0^{n_i^0} 1^{n_i^1}) ,$$

so, we will be able to rely on:

$$\frac{P^n_{\boldsymbol{\theta}}(\mathbf{x})}{Q^*(\mathbf{x})} = \frac{P^n_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{N})}{Q^*(\mathbf{x} \mid \mathbf{N})} .$$

Now, the average redundancy of Q^* with respect to $P^n_{\boldsymbol{\theta}}$ can be rewritten in a handy way.

$$\begin{aligned}
\mathbb{E}_W \left[D\left(P^n_{\boldsymbol{\theta}}, Q^*\right) \right] &= \mathbb{E}_W \left[\mathbb{E}_{P^n_{\boldsymbol{\theta}}} \left[\log \frac{P^n_{\boldsymbol{\theta}}(X_{1:n} \mid \mathbf{N})}{Q^*(X_{1:n} \mid \mathbf{N})} \right] \right] \\
&\quad \text{from the last equation,} \\
&= \mathbb{E}_W \left[\mathbb{E}_{P^n_{\boldsymbol{\theta}}} \left[\mathbb{E}_{P^n_{\boldsymbol{\theta}}} \left[\log \frac{P^n_{\boldsymbol{\theta}}(X_{1:n} \mid \mathbf{N})}{Q^*(X_{1:n} \mid \mathbf{N})} | \mathbf{N} \right] \right] \right] \\
&= \mathbb{E}_W \left[\mathbb{E}_{P^n_{\boldsymbol{\theta}}} \left[D\left(P^n_{\boldsymbol{\theta}}(\cdot \mid \mathbf{N}), Q^*_n(\cdot \mid \mathbf{N})\right) \right] \right] \\
&= \mathbb{E}_W \left[\mathbb{E}_{\mathbf{N}} \left[D\left(P^n_{\boldsymbol{\theta}}(\cdot \mid \mathbf{N}), Q^*(\cdot \mid \mathbf{N})\right) \right] \right] \\
&\quad \text{as the distribution of } \mathbf{N} \text{ does not depend on } \boldsymbol{\theta}, \\
&= \mathbb{E}_{\mathbf{N}} \left[\mathbb{E}_W \left[D\left(P^n_{\boldsymbol{\theta}}(\cdot \mid \mathbf{N}), Q^*(\cdot \mid \mathbf{N})\right) \right] \right] \\
&\quad \text{by Fubini's Theorem.}
\end{aligned}$$

We may develop $D\left(P_{\boldsymbol{\theta}}^n(\cdot \mid \mathbf{N}), Q^*(\cdot \mid \mathbf{N})\right)$ for a given value of $\mathbf{N} = (n_1, n_2, \dots, n_p)$. As both $P_{\boldsymbol{\theta}}^n(\cdot \mid \mathbf{N})$ and $Q^*(\cdot \mid \mathbf{N})$ are product distributions on $\prod_{i=1}^p \left(\{2i-1, 2i\}^{n_i}\right)$, we have

$$\begin{aligned}\mathbb{E}_W\left[D\left(P_{\boldsymbol{\theta}}^n(\cdot \mid \mathbf{N}), Q^*(\cdot \mid \mathbf{N})\right)\right] &= \mathbb{E}_W\left[\sum_{i=1}^p D\left(P_{\theta_i}^{n_i}, m_{n_i}^*\right)\right] \\ &= \sum_{i=1}^p \mathbb{E}_{W_i}\left[D\left(P_{\theta_i}^{n_i}, m_{n_i}^*\right)\right] .\end{aligned}$$

The minimal average redundancy of Λ_f^n with respect to the mixing distribution W is thus finally given by:

$$\begin{aligned}\mathbb{E}_W\left[D\left(P_{\boldsymbol{\theta}}^n, Q^*\right)\right] &= \mathbb{E}_{\mathbf{N}}\left[\sum_{i=1}^p \mathbb{E}_{W_i}\left[D\left(P_{\theta_i}^{n_i}, m_{n_i}^*\right)\right]\right] \\ &= \sum_{i=1}^p \mathbb{E}_{\mathbf{N}}\left[\mathbb{E}_{W_i}\left[D\left(P_{\theta_i}^{n_i}, m_{n_i}^*\right)\right]\right] \\ &= \sum_{i=1}^p \sum_{n_i=0}^n \binom{n}{n_i} \mu_i^{n_i}(1-\mu_i)^{n-n_i} \mathbb{E}_{W_i}\left[D\left(P_{\theta_i}^{n_i}, m_{n_i}^*\right)\right] . \end{aligned} \tag{1}$$

Hence, the minimal redundancy of Λ_f^n with respect to prior probability W is a weighted average of redundancies of Krichevsky-Trofimov mixtures over binary strings with different lengths.

At some place, we will use the Chebychef-Cantelli inequality (see Devroye et al., 1996) which asserts that for a square-integrable random variable:

$$\Pr\left\{X \leq \mathbb{E}[X] - t\right\} \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2} .$$

Besides, note that for all $\epsilon < \frac{1}{2}$,

$$\int_\epsilon^{1-\epsilon} \frac{\mathrm{d}x}{\pi\sqrt{x(1-x)}} > 1 - \frac{4}{\pi}\sqrt{\epsilon}. \tag{2}$$

Now, the proposition is derived by processing the right-hand-side of Equation (1).
Under condition $(1-\lambda)\, n\frac{f(2p)}{c(p)} > \frac{10}{\epsilon}$, we have $\frac{5}{n_i\epsilon} < \frac{1}{2}$ for all $i \le p$ such that $n_i \ge (1-\lambda)\, n\mu_i$. Hence,

$$
\begin{aligned}
&\mathbb{E}_W\left[D\left(P^n_{\boldsymbol{\theta}}, Q^*\right)\right] \\
&\ge \sum_{i=1}^{p} \sum_{n_i \ge (1-\lambda)n\mu_i}^{n} \binom{n}{n_i} \mu_i^{n_i}(1-\mu_i)^{n-n_i} \int_0^1 \frac{D\left(P^{n_i}_{\theta_i}, m^*_{n_i}\right)}{\sqrt{\theta_i(1-\theta_i)}}\, \mathrm{d}\theta_i \\
&\ge \sum_{i=1}^{p} \sum_{n_i \ge (1-\lambda)n\mu_i}^{n} \binom{n}{n_i} \mu_i^{n_i}(1-\mu_i)^{n-n_i} \int_{\frac{5}{n_i\,\epsilon}}^{1-\frac{5}{n_i\,\epsilon}} \frac{\frac{1}{2}\log\frac{n_i}{2\pi \mathrm{e}} + \log\pi - \epsilon}{\pi\sqrt{\theta_i(1-\theta_i)}}\, \mathrm{d}\theta_i \\
&\quad \text{by Proposition 1 from Xie and Barron (1997)} \\
&\ge \sum_{i=1}^{p} \sum_{n_i \ge (1-\lambda)n\mu_i}^{n} \binom{n}{n_i} \mu_i^{n_i}(1-\mu_i)^{n-n_i} \left(1 - \frac{4}{\pi}\sqrt{\frac{5}{n_i\epsilon}}\right)\left(\frac{1}{2}\log\frac{n_i}{2\pi\mathrm{e}} + \log\pi - \epsilon\right) \\
&\quad \text{from (2)} \\
&\ge \sum_{i=1}^{p} \sum_{n_i \ge (1-\lambda)n\mu_i}^{n} \binom{n}{n_i} \mu_i^{n_i}(1-\mu_i)^{n-n_i} \left(1 - \frac{4}{\pi}\sqrt{\frac{5}{(1-\lambda)\, n\mu_i\epsilon}}\right)\left(\frac{1}{2}\log\frac{n(1-\lambda)\mu_i}{2\pi\mathrm{e}} + \log\pi - \epsilon\right) \\
&\quad \text{using monotonicity of } x\log x \\
&\ge \sum_{i=1}^{p} \frac{1}{1+\frac{1-\mu_i}{n\,\mu_i\,\lambda^2}} \left(1 - \frac{4}{\pi}\sqrt{\frac{5}{(1-\lambda)\, n\mu_i\epsilon}}\right)\left(\frac{1}{2}\log\frac{n(1-\lambda)\mu_i}{2\pi\mathrm{e}} + \log\pi - \epsilon\right) \\
&\quad \text{invoking the Chebychef-Cantelli inequality,} \\
&\ge \frac{1}{1+\frac{c(p)}{\lambda^2 n f(2p)}} \left(1 - \frac{4}{\pi}\sqrt{\frac{5c(p)}{(1-\lambda)\, \epsilon n f(2p)}}\right) \sum_{i=1}^{p} \left(\frac{1}{2}\log\frac{n(1-\lambda)f(2i)}{2c(p)\pi\mathrm{e}} + \log\pi - \epsilon\right) \\
&\quad \text{using monotonicity assumption on } f.
\end{aligned}
$$

■

IV. Examples of envelope classes

Theorems 3, 4 and 5 assert that the summability of the envelope defining a class of memoryless sources characterizes the (strong) universal compressibility of that class. However, it is not easy to figure out whether the bounds provided by the last two theorems are close to each other or not. In this Section, we investigate the case of envelopes which decline either like power laws or exponentially fast. In both cases, upper-bounds on minimax regret will follow directly from Theorem 4 and a straightforward optimization. Specific lower bounds on minimax redundancies are derived by mimicking the proof of Theorem 5, either faithfully as in the case of exponential envelopes or by developing an alternative prior as in the case of power-law envelopes.

A. Power-law envelope classes

Let us first agree on the classical notation: $\zeta(\alpha) = \sum_{k\ge 1} \frac{1}{k^\alpha}$, for $\alpha > 1$.

Theorem 6: Let α denote a real number larger than 1, and C be such that $C\zeta(\alpha) \ge 2^\alpha$. The source class $\Lambda_{C\,\cdot^{-\alpha}}$ is the envelope class associated with the decreasing function $f_{\alpha,C} : x \mapsto 1 \wedge \frac{C}{x^\alpha}$ for $C > 1$ and $\alpha > 1$.
Then:

1)

$$n^{1/\alpha}\,A(\alpha)\,\log\left\lfloor (C\zeta(\alpha))^{1/\alpha}\right\rfloor \le R^+(\Lambda^n_{C\,\cdot^{-\alpha}})$$

where

$$A(\alpha) = \frac{1}{\alpha}\int_1^\infty \frac{1}{u^{1-1/\alpha}}\left(1-\mathrm{e}^{-1/(\zeta(\alpha)u)}\right)\mathrm{d}u\,.$$

2)

$$R^*(\Lambda^n_{C\,\cdot^{-\alpha}}) \le \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha}(\log n)^{1-1/\alpha} + O(1)\,.$$

Remark 8: The gap between the lower-bound and the upper-bound is of order $(\log n)^{1-\frac{1}{\alpha}}$. We are not in a position to claim that one of the two bounds is tight, let alone which one is tight. Note however that as $\alpha\to\infty$ and $C = H^\alpha$, class $\Lambda_{C\,\cdot^{-\alpha}}$ converges to the class of memoryless sources on alphabet $\{1,\ldots,H\}$ for which the minimax regret is $\frac{H-1}{2}\log n$. This is (up to a factor 2) what we obtain by taking the limits in our upper-bound of $R^*(\Lambda^n_{C\,\cdot^{-\alpha}})$. On the other side, the limit of our lower-bound when α goes to 1 is infinite, which is also satisfying since it agrees with Theorem 3.

Remark 9: In contrast with various lower bounds derived using a similar methodology, the proof given here relies on a single prior probability distribution on the parameter space and works for all values of n.

Note that the lower bound that can be derived from Theorem 5 is of the same order of magnitude $O(n^{1/\alpha})$ as the lower bound stated here (see Appendix II). The proof given here is completely elementary and does not rely on the subtle computations described in Xie and Barron (1997).

Proof: For the upper-bound on minimax regret, note that

$$\bar{F}_{\alpha,C}(u) = \sum_{k>u} 1\wedge\frac{C}{k^\alpha} \le \frac{C}{(\alpha-1)\,u^{\alpha-1}}.$$

Hence, choosing $u_n = \left(\frac{2Cn}{(\alpha-1)\log n}\right)^{\frac{1}{\alpha}}$, resorting to Theorem 4, we get:

$$R^*(\Lambda^n_{C\,\cdot^{-\alpha}}) \le \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha}(\log n)^{1-1/\alpha} + O(1).$$

Let us now turn to the lower bound. We first define a finite set Θ of parameters such that $P^n_{\boldsymbol\theta}\in\Lambda^n_{\alpha,C}$ for any $\boldsymbol\theta\in\Theta$ and then we use the mutual information lower bound on redundancy.

Let m be a positive integer such that $m^\alpha \le C\zeta(\alpha)$.

The set $\{P_{\boldsymbol\theta},\boldsymbol\theta\in\Theta\}$ consists of memoryless sources over the finite alphabet $\mathbb{N}_+$. Each parameter $\boldsymbol\theta$ is a sequence of integers $\boldsymbol\theta = (\theta_1,\theta_2,\ldots,)$. We take a prior distribution on Θ such that $(\theta_k)_k$ is a sequence of independent identically distributed random variables with uniform distribution on $\{1,\ldots,m\}$. For any such $\boldsymbol\theta$, $P^1_{\boldsymbol\theta}$ is a probability distribution on $\mathbb{N}_+$ with support $\cup_{k\ge1}\{(k-1)m+\theta_k\}$, namely:

$$P_{\boldsymbol\theta}((k-1)m+\theta_k) = \frac{1}{\zeta(\alpha)k^\alpha} = \frac{m^\alpha}{\zeta(\alpha)}\cdot\frac{1}{(k\,m)^\alpha}\quad\text{for } k\ge 1. \tag{3}$$

The condition $m^\alpha\le C\zeta(\alpha)$ ensures that $P^1_{\boldsymbol\theta}\in\Lambda^1_{\alpha,C}$.

Now, the mutual information between parameter $\boldsymbol{\theta}$ and source output $X_{1:n}$ is

$$I\left(\boldsymbol{\theta}, X_{1:n}\right) = \sum_{k \geq 1} I\left(\theta_k, X_{1:n}\right)$$

Let $N_k\left(\mathbf{x}\right) = 1$ if there exists some index $i \in \{1, \ldots, n\}$ such that $\mathbf{x}_i \in [(k-1)m+1, km]$, and 0 otherwise. Note that the distribution of N_k does not depend on the value of $\boldsymbol{\theta}$. Thus we can write:

$$I\left(\theta_k, X_{1:n}\right) = I\left(\theta_k, X_{1:n} | N_k = 0\right) P\left(N_k = 0\right) + I\left(\theta_k, X_{1:n} | N_k = 1\right) P\left(N_k = 1\right).$$

But, conditionally on $N_k = 0$, θ_k and $X_{1:n}$ are independent. Moreover, conditionally on $N_k = 1$ we have

$$I\left(\theta_k, X_{1:n} \mid N_k = 1\right) = \mathbb{E}\left[\log \frac{P(\theta_k = j | X_{1:n})}{P(\theta_k = j)} | N_k = 1\right] = \log m.$$

Hence,

$$\begin{aligned} I\left(\boldsymbol{\theta}, X_{1:n}\right) &= \sum_{k \geq 1} P(N_k = 1) \log m \\ &= \mathbb{E}_{P_{\boldsymbol{\theta}}}\left[Z_n\right] \log m, \end{aligned}$$

where $Z_n(\mathbf{x})$ denotes the number of distinct symbols in string $\mathbf{x}$ (note that its distribution does not depend on the value of $\boldsymbol{\theta}$.) As $Z_n = \sum_{k \geq 1} 1_{N_k(x)=1}$, the expectation translates into a sum

$$\mathbb{E}_{\theta}\left[Z_n\right] = \sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{1}{\zeta(\alpha) k^{\alpha}}\right)^n\right)$$

which leads to:

$$R^{+}(\Lambda^n_{\alpha,C}) \geq \left(\sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{1}{\zeta(\alpha) k^{\alpha}}\right)^n\right)\right) \times \log m\,.$$

Now:

$$\begin{aligned} &\sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{1}{\zeta(\alpha) k^{\alpha}}\right)^n\right) \\ &\geq \sum_{k=1}^{\infty} \left(1 - \exp\left(-\frac{n}{\zeta(\alpha)\, k^{\alpha}}\right)\right) \\ &\quad \text{as } 1 - x \leq \exp(-x) \\ &\geq \int_1^{\infty} \left(1 - \exp\left(-\frac{n}{\zeta(\alpha) x^{\alpha}}\right)\right) \mathrm{d}x \\ &\geq \frac{n^{\frac{1}{\alpha}}}{\alpha} \int_1^{\infty} \frac{1}{u^{1-\frac{1}{\alpha}}} \left(1 - \exp\left(-\frac{1}{\zeta(\alpha) u}\right)\right) \mathrm{d}u\,. \end{aligned}$$

■

In order to optimize the bound we choose the largest possible m which is $m = \lfloor (C\zeta(\alpha))^{1/\alpha} \rfloor$.

For an alternative derivation of a similar lower-bound using Theorem 5, see Appendix II.

B. Exponential envelope classes

Theorems 4 and 5 provide almost matching bounds on the minimax redundancy for source classes defined by exponentially vanishing envelopes.

Theorem 7: Let C and α denote positive real numbers satisfying $C > e^{2\alpha}$. The class $\Lambda_{Ce^{-\alpha\cdot}}$ is the envelope class associated with function $f_\alpha : x \mapsto 1 \wedge C\mathrm{e}^{-\alpha x}$. Then

$$\frac{1}{8\alpha}\log^2 n\,(1-o(1)) \le R^+(\Lambda^n_{Ce^{-\alpha\cdot}}) \le R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \le \frac{1}{2\alpha}\log^2 n + O(1)$$

Proof: For the upper-bound, note that

$$\bar{F}_\alpha(u) = \sum_{k>u} 1 \wedge C\mathrm{e}^{-\alpha k} \le \frac{C}{1-\mathrm{e}^{-\alpha}}\,\mathrm{e}^{-\alpha(u+1)}\,.$$

Hence, by choosing the optimal value $u_n = \frac{1}{\alpha}\log n$ in Theorem 4 we get:

$$R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \le \frac{1}{2\alpha}\log^2 n + O(1).$$

We will now prove the lower bound using Theorem 5. The constraint $C > \mathrm{e}^{2\alpha}$ warrants that the sequence $c(p) = \sum_{k=1}^{p} f(2k) \ge C\mathrm{e}^{-2\alpha}\frac{1-\mathrm{e}^{-2\alpha p}}{1-\mathrm{e}^{-2\alpha}}$ is larger than 1 for all p.

If we choose $p = \left\lfloor \frac{1}{2\alpha}(\log n - \log\log n)\right\rfloor$, then $nf(2p) > Cne^{-\log n + \log\log n - 2\alpha}$ goes to infinity with n. For $\epsilon = \lambda = \frac{1}{2}$, we get $C(p,n,\lambda,\epsilon) = 1 - o(1)$. Besides,

$$\begin{aligned}
\sum_{i=1}^{p}\left(\frac{1}{2}\log\frac{n(1-\lambda)C\pi e^{-2\alpha i}}{2c(p)e} - \epsilon\right) &= \frac{p}{2}\left(\log n + \log\frac{(1-\lambda)C\pi}{2c(p)e} - 2\epsilon\right) - \alpha\sum_{i=1}^{p} i \\
&= \left(\frac{1}{4\alpha}\log^2 n - \frac{\alpha}{2}\frac{1}{4\alpha^2}\log^2 n\right)(1+o(1)) \\
&= \frac{1}{8\alpha}\log^2 n\,(1+o(1))\,.
\end{aligned}$$

■

V. A CENSORING CODE FOR ENVELOPE CLASSES

The proof of Theorem 4 suggests to handle separately small and large (allegedly infrequent) symbols. Such an algorithm should perform quite well as soon as the tail behavior of the envelope provides an adequate description of the sources in the class. The coding algorithm suggested by the proof of Theorem 4, which are based on the Shtarkov NML coder, are not computationally attractive. The design of the next algorithm (`CensoringCode`) is again guided by the proof of Theorem 4: it is parameterized by a sequence of cutoffs $(K_i)_{i\in\mathbb{N}}$ and handles the i^{th} symbol of the sequence to be encoded differently according to whether it is smaller or larger than cutoff K_i, in the latter situation, the symbol is said to be censored. The `CensoringCode` algorithm uses Elias penultimate code (Elias, 1975) to encode censored symbols and Krichevsky-Trofimov mixtures (Krichevsky and Trofimov, 1981) to encode the sequence of non-censored symbols padded with markers (zeros) to witness acts of censorship. The performance of this algorithm is evaluated on the power-law envelope class $\Lambda_{C\cdot^{-\alpha}}$, already investigated in Section IV. In this

section, the parameters α and C are assumed to be known.

Let us first describe the algorithm more precisely. Given a non-decreasing sequence of cutoffs $(K_i)_{i\leq n}$, a string $\mathbf{x}$ from $\mathbb{N}_+^n$ defines two strings $\tilde{\mathbf{x}}$ and $\check{\mathbf{x}}$ in the following way. The i^{th} symbol $\mathbf{x}_i$ of $\mathbf{x}$ is censored if $\mathbf{x}_i > k_i$. String $\tilde{\mathbf{x}}$ has length n and belongs to $\prod_{i=1}^n \mathcal{X}_i$, where $\mathcal{X}_i = \{0, \ldots K_i\}$:

$$\tilde{\mathbf{x}}_i = \begin{cases} \mathbf{x}_i \text{ if } \mathbf{x}_i \leq K_i \\ 0 \text{ otherwise (the symbol is censored).} \end{cases}$$

Symbol 0 serves as an escape symbol. Meanwhile, string $\check{\mathbf{x}}$ is the subsequence of censored symbols, that is $(\mathbf{x}_i)_{\mathbf{x}_i > K_i, i\leq n}$.

The algorithm encodes $\mathbf{x}$ as a pair of binary strings `C1` and `C2`. The first one (`C1`) is obtained by applying Elias penultimate code to each symbol from $\check{\mathbf{x}}$, that is to each censored symbol. The second string (`C2`) is built by applying arithmetic coding to $\tilde{\mathbf{x}}$ using side-information from $\check{\mathbf{x}}$. Decoding `C2` can be sequentially carried out using information obtained from decoding `C1`.

In order to describe the coding probability used to encode $\tilde{\mathbf{x}}$, we need a few more counters. For $j > 0$, let n_i^j be the number of occurrences of symbol j in $\mathbf{x}_{1:i}$ and let n_i^0 be the number of symbols larger than K_{i+1} in $\mathbf{x}_{1:i}$ (note that this not larger than the number of censored symbols in $\mathbf{x}_{1:i}$, and that the counters $n_i^{\cdot}$ can be recovered from $\tilde{\mathbf{x}}_{1:i}$ and $\check{\mathbf{x}}$). The conditional coding probability over alphabet $\mathcal{X}_{i+1} = \{0, \ldots, K_{i+1}\}$ given $\tilde{\mathbf{x}}_{1:i}$ and $\check{\mathbf{x}}$ is derived from the Krichevsky-Trofimov mixture over $\mathcal{X}_{i+1}$. It is the posterior distribution corresponding to Jeffrey's prior on the $1 + K_{i+1}$-dimensional probability simplex and counts n_i^j for j running from 0 to K_{i+1}:

$$Q(\tilde{X}_{i+1} = j \mid \tilde{X}_{1:i} = \tilde{\mathbf{x}}_{1:i}, \check{X} = \check{\mathbf{x}}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{K_{i+1}+1}{2}} \, .$$

The length of $\mathtt{C2}(\mathbf{x})$ is (up to a quantity smaller than 1) given by

$$-\sum_{i=0}^{n-1} \log Q(\tilde{\mathbf{x}}_{i+1} \mid \mathbf{x}_{1:i}, \check{\mathbf{x}}) = -\log Q(\tilde{\mathbf{x}} \mid \check{\mathbf{x}}) \, .$$

The following description of the coding probability will prove useful when upper-bounding redundancy. For $1 \leq j \leq K_n$, let s^j be the number of censored occurrences of symbol j. Let n^j serve as a shorthand for n_n^j. Let $i(j)$ be n if $K_n < j$ or the largest integer i such that K_i is smaller than j, then $s_j = n_{i(j)}^j$. The following holds

$$Q(\tilde{\mathbf{x}} \mid \check{\mathbf{x}}) = \left(\prod_{j=1}^{K_n} \frac{\Gamma(n^j + 1/2)}{\Gamma(s^j + 1/2)} \right) \left(\prod_{i:\tilde{\mathbf{x}}_i = 0} (n_{i-1}^0 + 1/2) \right) \left(\prod_{i=0}^{n-1} \frac{1}{i + \frac{K_{i+1}+1}{2}} \right)$$

Note that the sequence $(n_i^0)_{i\leq n}$ is not necessarily non-decreasing.

A technical description of algorithm `CensoringCode` is given below. The procedure `EliasCode` takes as input an integer j and outputs a binary encoding of j using exactly $\ell(j)$ bits where ℓ is defined by: $\ell(j) = \lfloor \log j + 2 \log(1 + \log j) + 1 \rfloor$. The procedure `ArithCode` builds on the arithmetic coding methodology (Rissanen and Langdon, 1979). It is enough to remember that an arithmetic coder takes advantage of the fact that a coding probability Q is

completely defined by the sequence of conditional distributions of the $i+1$th symbol given the past up to time i.

The proof of the upper-bound in Theorem 6, prompts us to choose $K_i = \lambda i^{\frac{1}{\alpha}}$, it will become clear afterward that a reasonable choice is $\lambda = \left(\frac{4C}{\alpha-1}\right)^{\frac{1}{\alpha}}$.

Algorithm 1 CensoringCode

```
K ← 0
counts ← [1/2, 1/2, ...]
for i from 1 to n do
  cutoff ← ⌊(4 Ci/(α−1))^{1/α}⌋
  if cutoff > K then
    for j ← K + 1 to cutoff do
      counts[0] ← counts[0] − counts[j] + 1/2
    end for
    K ← cutoff
  end if
  if x[i] ≤ cutoff then
    ArithCode(x[i], counts[0 : cutoff])
  else
    ArithCode(0, counts[0 : cutoff])
    C1←C1·EliasCode(x[i])
    counts[0] ← counts[0] + 1
  end if
  counts[x[i]] ← counts[x[i]] + 1
end for
C2← ArithCode()
C1 · C2
```

Theorem 8: Let C and α be positive reals. Let the sequence of cutoffs $(K_i)_{i\leq n}$ be given by

$$K_i = \left\lfloor \left(\frac{4Ci}{\alpha-1}\right)^{1/\alpha} \right\rfloor .$$

The expected redundancy of procedure `CensoringCode` on the envelope class $\Lambda_{C\cdot^{-\alpha}}$ is not larger than

$$\left(\frac{4Cn}{\alpha-1}\right)^{\frac{1}{\alpha}} \log n \left(1+o(1)\right) .$$

Remark 10: The redundancy upper-bound in this Theorem is within a factor $\log n$ from the lower bound $O(n^{1/\alpha})$ from Theorem 6 .

The proof of the Theorem builds on the next two lemmas. The first lemma compares the length of $\mathtt{C2}(\mathbf{x})$ with a tractable quantity. The second lemma upper-bounds the average length of $\mathtt{C1}(\mathbf{x})$ by a quantity which is of the same order of magnitude as the upper-bound on redundancy we are looking for.

We need a few more definitions. Let $\mathbf{y}$ be the string of length n over alphabet $\mathcal{X}_n$ defined by:

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_i \text{ if } \mathbf{x}_i \leq K_n; \\ 0 \text{ else.} \end{cases}$$

For $0 \leq j \leq K_n$, note that the previously defined shorthand n^j is the number of occurrences of symbol j in $\mathbf{y}$.

The string $\mathbf{y}$ is obtained from $\mathbf{x}$ in the same way as $\tilde{\mathbf{x}}$ using the constant cutoff K_n.

Let m_n^* be the Krichevsky-Trofimov mixture over alphabet $\{0, \ldots, K_n\}$:

$$m_n^*(\mathbf{y}) = \left(\prod_{j=0}^{K_n} \frac{\Gamma(n^j + \frac{1}{2})}{\Gamma(1/2)} \right) \frac{\Gamma(\frac{K_n+1}{2})}{\Gamma\left(n + \frac{K_n+1}{2}\right)} .$$

String $\tilde{\mathbf{x}}$ seems easier to encode than $\mathbf{y}$ since it is possible to recover $\tilde{\mathbf{x}}$ from $\mathbf{y}$. This observation does not however warrant automatically that the length of $\texttt{C2}(\mathbf{x})$ is not significantly larger than any reasonable codeword length for $\mathbf{y}$. Such a guarantee is provided by the following lemma.

Lemma 2: For every string $\mathbf{x} \in \mathbb{N}_+^n$, the length of the $\texttt{C2}(\mathbf{x})$ is not larger than $-\log m_n^*(\mathbf{y})$.

Proof: [Lemma 2] Let s^0 be the number of occurrences of 0 in $\mathbf{y}$, that is the number of symbols in $\mathbf{x}$ that are larger than K_n. Let

$$T_0 = \prod_{i=1, \tilde{\mathbf{x}}_i = 0}^{n} (n_{i-1}^0 + 1/2) .$$

Then, the following holds:

$$\begin{aligned} T_0 &\overset{(a)}{=} \left(\prod_{i=1, \mathbf{y}_i = 0}^{n} (n_{i-1}^0 + 1/2) \right) \prod_{j=1}^{K_n} \left(\prod_{i=1, \mathbf{y}_i = j}^{i(j)} (n_{i-1}^0 + 1/2) \right) \\ &\overset{(b)}{\geq} \left(\frac{\Gamma(s^0 + 1/2)}{\Gamma(1/2)} \right) \prod_{j=1}^{K_n} \left(\frac{\Gamma(s^j + 1/2)}{\Gamma(1/2)} \right) , \end{aligned}$$

where (a) follows from the fact symbol $\mathbf{x}_i$ is censored either because $\mathbf{x}_i > K_n$ (that is $\mathbf{y}_i = 0$) or because $\mathbf{x}_i = j \leq K_n$ and $i \leq i(j)$; (b) follows from the fact that for each $i \leq n$ such that $\mathbf{y}_i = 0$, $n_{i-1}^0 \geq \sum_{i' < i} \mathbf{1}_{\mathbf{x}_{i'} > K_n}$ while for each $j, 0 < j \leq K_n$, for each $i \leq i(j)$, $n_{i-1}^0 \geq n_{i-1}^j$.

From the last inequality, it follows that

$$\begin{aligned} Q(\tilde{\mathbf{x}} \mid \check{\mathbf{x}}) &\geq \left(\prod_{j=1}^{K_n} \frac{\Gamma(n^j + 1/2)}{\Gamma(s^j + 1/2)} \right) \frac{\Gamma(s^0 + 1/2)}{\Gamma(1/2)} \prod_{j=1}^{K_n} \frac{\Gamma(s^j + 1/2)}{\Gamma(1/2)} \left(\prod_{i=0}^{n-1} \frac{1}{i + \frac{K_{i+1}+1}{2}} \right) \\ &\geq \left(\prod_{j=1}^{K_n} \frac{\Gamma(n^j + 1/2)}{\Gamma(s^j + 1/2)} \right) \frac{\Gamma(s^0 + 1/2)}{\Gamma(1/2)} \prod_{j=1}^{K_n} \frac{\Gamma(s^j + 1/2)}{\Gamma(1/2)} \left(\prod_{i=0}^{n-1} \frac{1}{i + \frac{K_n+1}{2}} \right) \\ &= m_n^*(\mathbf{y}) , \end{aligned}$$

where the last inequality holds since $(K_i)_i$ is a non-decreasing sequence. ■

The next lemma shows that the expected length of $\texttt{C1}(X_{1:n})$ is not larger than the upper-bound we are looking for.

Lemma 3: For every source $P \in \Lambda_{C \cdot^{-\alpha}}$, the expected length of the encoding of the censored symbols ($\texttt{C1}(X_{1:n})$) satisfies:

$$\mathbb{E}_P \left[\left| \texttt{C1}\left(X_{1:n}\right) \right| \right] \leq \frac{2C}{(\alpha - 1)\, \lambda^{\alpha-1}} n^{\frac{1}{\alpha}} \log n \,(1 + o(1)) .$$

Proof: [Lemma 3] Let $1 \le a < b$ and $\beta > 0$ but $\beta \neq 1$. Recall that we use binary logarithms and note that:

$$\int_a^b \frac{1}{x^\beta} \mathrm{d}x = \left[\frac{1}{(1-\beta)\, x^{\beta-1}}\right]_a^b, \tag{4}$$

$$\int_a^b \frac{\log x}{x^\beta} \mathrm{d}x = \left[\frac{\log x - \frac{\log e}{1-\beta}}{(1-\beta)\, x^{\beta-1}}\right]_a^b. \tag{5}$$

The expected length of the Elias encoding of censored symbols ($\mathtt{C1}(X_1^n)$)is:

$$\begin{aligned}
\mathbb{E}\big[|\mathtt{C1}\,(X_1^n)\,|\big] &= \mathbb{E}\left[\sum_{j=1}^n \ell(X_j)\mathbb{1}_{X_j > K_j}\right] \\
&= \sum_{j=1}^n \sum_{x=K_j+1}^{\infty} \ell(x)P(x) \\
&\le \sum_{j=1}^n \sum_{x=K_j+1}^{\infty} \ell(x)\frac{C}{x^\alpha} \\
&\le C \sum_{j=1}^n \sum_{x=K_j+1}^{\infty} \frac{\log(x) + 2\log\left(1+\log x\right) + 1}{x^\alpha}.
\end{aligned}$$

Note that for $x \ge 2^7$

$$\log x \ge 2\log(1+\log x) + 1,$$

so that the last sum is upper-bounded by

$$C \sum_{j=1}^n \sum_{x=K_j+1}^{\infty} 2\frac{\log x}{x^\alpha} + C \sum_{j:K_j<2^7} \sum_{x=1}^{2^7} \frac{2\log(1+\log x) + 1 - \log x}{x^\alpha}\,.$$

Using the expressions for the integrals above, we get:

$$\begin{aligned}
\sum_{x=K_j+1}^{\infty} \frac{\log x}{x^\alpha} &\le \int_{K_j}^{\infty} \frac{\log x}{x^\alpha} \mathrm{d}x \\
&\le \frac{\log K_j + \frac{\log e}{\alpha - 1}}{(\alpha-1)\, K_j^{\alpha-1}}.
\end{aligned}$$

Thus, as $K_j = \lambda j^{1/\alpha}$, let us denote by D_λ the expression

$$\sum_{j:j<(2^7/\lambda)^\alpha} \sum_{x=1}^{2^7} \frac{2\log(1+\log x) + 1 - \log x}{x^\alpha}\,.$$

Now, we substitute β by $1-\frac{1}{\alpha}$ in Equations (4) and (5) to obtain:

$$\begin{aligned}
\mathbb{E}\big[|\mathtt{C1}\,(X_1^n)\,|\big] &\leq CD_\lambda + \frac{2C}{\alpha-1}\sum_{j=1}^{n}\frac{\log K_j + \frac{\log e}{\alpha-1}}{K_j^{\alpha-1}} \\
&\leq CD_\lambda + \frac{2C}{\alpha-1}\sum_{j=1}^{n}\frac{\frac{1}{\alpha}\log j + \log\lambda + \frac{\log e}{\alpha-1}}{\lambda^{\alpha-1}j^{1-\frac{1}{\alpha}}} \\
&\leq CD_\lambda + \frac{2C}{\alpha(\alpha-1)\lambda^{\alpha-1}}\left(C + \int_{x=2}^{n+1}\frac{\left(\log x + \alpha\log\lambda + \frac{\alpha\log e}{\alpha-1}\right)}{x^{1-\frac{1}{\alpha}}}\mathrm{d}x\right) \\
&= \frac{2C}{(\alpha-1)\,\lambda^{\alpha-1}}n^{\frac{1}{\alpha}}\log n\,(1+o(1))\,.
\end{aligned}$$

■

We may now complete the proof of Theorem 8.

Proof: Remember that $\mathcal{X}_n = \{0,\ldots,K_n\}$. If p is a probability mass function over alphabet $\mathcal{X}$, let $p^{\otimes n}$ be the probability mass function over $\mathcal{X}^n$ defined by $p^{\otimes n}(\mathbf{x}) = \prod_{i=1}^n p(\mathbf{x}_i)$. Note that for every string $\mathbf{x}\in\mathbb{N}_+^n$,

$$\max_{p\in\mathfrak{M}_1(\mathcal{X}_n)} p^{\otimes n}(\mathbf{y}) \geq \max_{p\in\mathfrak{M}_1(\mathbb{N}_+)} p^{\otimes n}(\mathbf{x}) \geq \max_{P\in\Lambda_{C\cdot-\alpha}} P^n(\mathbf{x}) = \hat{p}(\mathbf{x})\,.$$

Together with Lemma 2 and the bounds on the redundancy of the Krichevsky-Trofimov mixture (See Krichevsky and Trofimov, 1981), this implies:

$$|\mathtt{C2}\,(\mathbf{x})\,| \leq -\log\hat{p}\,(\mathbf{x}) + \frac{K_n}{2}\log n + O(1).$$

Let $L(\mathbf{x})$ be the length of the code produced by algorithm `CensoringCode` on the input string $\mathbf{x}$, then

$$\begin{aligned}
&\sup_{P\in\Lambda_{C\cdot-\alpha}} \mathbb{E}_P\left[L(X_{1:n}) - \log 1/P^n(X_{1:n})\right] \\
&\leq \sup_{P\in\Lambda_{C\cdot-\alpha}} \mathbb{E}_P\left[L(X_{1:n}) - \log 1/\hat{p}(X_{1:n})\right] \\
&\leq \sup_{P\in\Lambda^n_{C\cdot-\alpha}} \mathbb{E}_P\left[|\mathtt{C2}\,(X_{1:n})| + \log\hat{p}\,(X_{1:n}) + |\mathtt{C1}\,(X_{1:n})|\right] \\
&\leq \sup_{\mathbf{x}}\left(|\mathtt{C2}\,(\mathbf{x})| + \log\hat{p}\,(\mathbf{x})\right) + \sup_{P\in\Lambda^n_{C\cdot-\alpha}} \mathbb{E}_P\left[|\mathtt{C1}\,(X_{1:n})|\right] \\
&\leq \frac{\lambda n^{\frac{1}{\alpha}}}{2}\log n + \frac{2C}{(\alpha-1)\,\lambda^{\alpha-1}}n^{\frac{1}{\alpha}}\log n\,(1+o(1))\,.
\end{aligned}$$

The optimal value is $\lambda = \left(\frac{4C}{\alpha-1}\right)^{\frac{1}{\alpha}}$, for which we get:

$$R^+(Q^n,\Lambda^n_{C\cdot-\alpha}) \leq \left(\frac{4Cn}{\alpha-1}\right)^{\frac{1}{\alpha}}\log n\,(1+o(1))\,.$$

■

VI. ADAPTIVE ALGORITHMS

The performance of `CensoringCode` depends on the fit of the cutoffs sequence to the tail behavior of the envelope. From the proof of Theorem 8, it should be clear that if `CensoringCode` is fed with a source which marginal is light-tailed, it will be unable to take advantage of this, and will suffer from excessive redundancy.

In this section, a sequence $(Q^n)_n$ of coding probabilities is said to be *approximately asymptotically adaptive* with respect to a collection $(\Lambda_m)_{m\in\mathcal{M}}$ of source classes if for each $P \in \cup_{m\in\mathcal{M}}\Lambda_m$, for each Λ_m such that $P \in \Lambda_m$:

$$D(P^n, Q^n)/R^+(\Lambda_m^n) \in O(\log n)\,.$$

Such a definition makes sense, since we are considering massive source classes which minimax redundancies are large but still $o(\frac{n}{\log n})$. If each class Λ_m admits a non-trivial redundancy rate such that $R^+(\Lambda_m^n) = o(\frac{n}{\log n})$, the existence of an approximately asymptotically adaptive sequence of coding probabilities means that $\cup_m\Lambda_m$ is feebly universal (see the Introduction for a definition).

A. Pattern coding

First, the use of *pattern coding* Orlitsky et al. (2004), Shamir (2006) leads to an almost minimax adaptive procedure for small values of α, that is heavy-tailed distributions. Let us introduce the notion of pattern using the example of string $\mathbf{x} =$ “abracadabra”, which is made of $n = 11$ characters. The information it conveys can be separated in two blocks:

1) a *dictionary* $\Delta = \Delta(\mathbf{x})$: the sequence of distinct symbols occurring in $\mathbf{x}$ in order of appearance (in the example, $\Delta = (a, b, r, c, d)$).
2) a *pattern* $\psi = \psi(\mathbf{x})$ where ψ_i is the rank of $\mathbf{x}_i$ in the dictionary Δ (here, $\psi = 1231415123$).

Now, consider the algorithm coding message $\mathbf{x}$ by transmitting successively:

1) the dictionary $\Delta_n = \Delta\left(\mathbf{x}\right)$ (by concatenating the Elias codes for successive symbols);
2) and the pattern $\Psi_n = \psi\left(\mathbf{x}\right)$, using a minimax procedure for coding patterns as suggested by Orlitsky et al. (2004) or Shamir (2006). Henceforth, the latter procedure is called pattern coding.

Theorem 9: Let Q^n denote the coding probability associated with the coding algorithm which consists in applying Elias penultimate coding to the dictionary $\Delta(\mathbf{x})$ of a string $\mathbf{x}$ from $\mathbb{N}_+^n$ and then pattern coding to the pattern $\psi(\mathbf{x})$.

Then for any α such that $1 < \alpha \leq 5/2$, there exists a constant K depending on α and C such that

$$R^+(Q^n, \Lambda^n_{C\cdot^{-\alpha}}) \leq K n^{1/\alpha}\log n$$

Proof: For a given value of C and α, the Elias encoding of the dictionary uses on average

$$\mathbb{E}\left[|\Delta_n|\right] = K' n^{\frac{1}{\alpha}} \log n$$

bits (as proved in Appendix IV), for some constant K' depending on α and C.

If our pattern coder reaches (approximately) the minimax pattern redundancy

$$R^+_{\Psi}\left(\Psi_{1:n}\right) = \inf_{q \in \mathfrak{M}_1\left(\mathbb{N}_+^n\right)} \sup_{P \in \mathfrak{M}_1(\mathbb{N}_+)} \mathbb{E}_P \left[\log \frac{P^{\otimes n}(\Psi_{1:n})}{q(\Psi_{1:n})} \right],$$

the encoding of the pattern uses on average

$$H(\Psi_{1:n}) + R^+_{\Psi}\left(\Psi_{1:n}\right) \leq H(X_{1:n}) + R^+_{\Psi}\left(\Psi_{1:n}\right) \text{ bits.}$$

But in Orlitsky et al. (2004), the authors show that $R^+_{\Psi}\left(\Psi_{1:n}\right)$ is upper-bounded by $O\left(\sqrt{n}\right)$ and even $O\left(n^{\frac{2}{5}}\right)$ according to Shamir (2004) (actually, these bounds are even satisfied by the minimax individual pattern redundancy). ■

This remarkably simple method is however expected to have a poor performance when α is large. Indeed, it is proved in Garivier (2006) that $R^+_{\Psi}\left(\Psi_{1:n}\right)$ is lower-bounded by $1.84\left(\frac{n}{\log n}\right)^{\frac{1}{3}}$ (see also Shamir (2006) and references therein), which indicates that pattern coding is probably suboptimal as soon as α is larger than 3.

B. An approximately asymptotically adaptive censoring code

Given the limited scope of the pattern coding method, we will attempt to turn the censoring code into an adaptive method, that is to tune the cutoff sequence so as to model the source statistics. As the cutoffs are chosen in such a way that they model the tail-heaviness of the source, we are facing a tail-heaviness estimation problem . In order to focus on the most important issues we do not attempt to develop a sequential algorithm. The $n+1$th cutoff K_{n+1} is chosen according to the number of *distinct* symbols $Z_n(\mathbf{x})$ in $\mathbf{x}$.

This is a reasonable method if the probability mass function defining the source statistics P^1 actually decays like $\frac{1}{k^\alpha}$. Unfortunately, sparse distributions consistent with $\Lambda_{\cdot^{-\alpha}}$ may lead this project astray. If, for example, $(Y_n)_n$ is a sequence of geometrically distributed random variables, and if $X_n = \left\lfloor 2^{\frac{Y_n}{\alpha}} \right\rfloor$, then the distribution of the X_n just fits in $\Lambda_{C\cdot^{-\alpha}}$ but obviously $Z_n(X_{1:n}) = Z_n\left(Y_{1:n}\right) = O\left(\log n\right)$.

Thus, rather than attempting to handle $\cup_{\alpha>0}\Lambda_{\cdot^{-\alpha}}$, we focus on subclasses $\cup_{\alpha>0}\mathcal{W}_\alpha$, where

$$\mathcal{W}_\alpha = \left\{ P \ : \ P \in \Lambda_{\cdot^{-\alpha}}, \ 0 < \liminf_k k^\alpha P^1(k) \leq \limsup_k k^\alpha P^1(k) < \infty \right\}.$$

The rationale for tuning cutoff K_n using Z_n comes from the following two propositions.

Proposition 7: For every memoryless source $P \in \mathcal{W}_\alpha$, there exist constants c_1 and c_2 such that for all positive integer n,

$$c_1 n^{1/\alpha} \leq \mathbb{E}[Z_n] \leq c_2 n^{1/\alpha}.$$

Proposition 8: The number of distinct symbols Z_n output by a memoryless source satisfies a Bernstein inequality:

$$P\left\{ Z_n \leq \frac{1}{2}\mathbb{E}[Z_n] \right\} \leq \mathrm{e}^{-\frac{\mathbb{E}[Z_n]}{8}}. \tag{6}$$

Proof: Note that Z_n is a function of n independent random variables. Moreover, Z_n is a configuration function as defined defined by Talagrand (1995) since $Z_n(\boldsymbol{x})$ is the size of a maximum subsequence of $\boldsymbol{x}$ satisfying

an hereditary property (all its symbols are pairwise distinct). Using the main theorem in Boucheron et al. (2000), this is enough to conclude. ■

Noting that $Z_n \geq 1$, we can derive the following inequality that will prove useful later on:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{Z_n^{\alpha-1}}\right] &= \mathbb{E}\left[\frac{1}{Z_n^{\alpha-1}}\mathbb{1}_{Z_n>\frac{1}{2}\mathbb{E}[Z_n]}\right] + \mathbb{E}\left[\frac{1}{Z_n^{\alpha-1}}\mathbb{1}_{Z_n\leq\frac{1}{2}\mathbb{E}[Z_n]}\right] \\ &\leq \frac{1}{\left(\frac{1}{2}\mathbb{E}[Z_n]\right)^{\alpha-1}} + P\left(Z_n \leq \frac{1}{2}\mathbb{E}[Z_n]\right). \end{aligned} \tag{7}$$

We consider here a modified version of `CensoringCode` that operates similarly, except that

1) the string $\mathbf{x}$ is first scanned completely to determine $Z_n(\mathbf{x})$;
2) the constant cutoff $\hat{K}_n = \mu Z_n$ is used for all symbols $\mathbf{x}_i$, $1 \leq i \leq n$, where μ is some positive constant.
3) the value of K_n is encoded using Elias penultimate code and transmitted before `C1` and `C2`.

Note that this version of the algorithm is not sequential because of the initial scanning.

Algorithm 2 AdaptiveCensoringCode

cutoff $\leftarrow \mu\, Z_n(\mathbf{x})$ {Determination of the constant cutoff}
counts $\leftarrow [1/2, 1/2, \ldots]$
for i from 1 to n **do**
  if $x[i] \leq$ *cutoff* **then**
    ArithCode($x[i]$, *counts*$[0$: *cutoff*$]$)
  else
    ArithCode(0, *counts*$[0$: *cutoff*$]$)
    `C1`←`C1`·EliasCode($x[i]$)
    counts$[0] \leftarrow$ *counts*$[0] + 1$
  end if
  counts$[x[i]] \leftarrow$ *counts*$[x[i]] + 1$
end for
`C2`← ArithCode()
$\mathtt{C}_1 \cdot \mathtt{C}_2$

We may now assert.

Theorem 10: The algorithm `AdaptiveCensoringCode` is approximately asymptotically adaptive with respect to $\bigcup_{\alpha>0} \mathcal{W}_\alpha$.

Proof: Let us again denote by $\mathtt{C1}(\mathbf{x})$ and $\mathtt{C2}(\mathbf{x})$ the two parts of the code-string associated with $\mathbf{x}$.

Let $\hat{L}$ be the codelength of the output of algorithm `AdaptiveCensoringCode`.

For any source P:

$$\begin{aligned}\mathbb{E}_P\left[\hat{L}\left(X_{1:n}\right)\right] - H(X_{1:n}) &= \mathbb{E}_P\left[\ell(\hat{K}_n) + |\mathtt{C1}\left(X_{1:n}\right)| + |\mathtt{C2}\left(X_{1:n}\right)|\right] - n\sum_{k=1}^{\infty} P^1(k)\log\frac{1}{P_1(k)} \\ &\leq \mathbb{E}_P\left[\ell(\hat{K}_n)\right] + \mathbb{E}_P\left[|\mathtt{C1}\left(X_{1:n}\right)|\right] + \mathbb{E}_P\left[|\mathtt{C2}\left(X_{1:n}\right)| - n\sum_{k=1}^{\hat{K}_n} P^1(k)\log\frac{1}{P_1(k)}\right].\end{aligned}$$

As function ℓ is increasing and equivalent to $\log$ at infinity, the first summand is obviously $o\left(\mathbb{E}_P\left[\ell(\hat{K}_n)\right]\right)$.

Moreover, if $P \in \mathcal{W}_\alpha$ there exists C such that $P^1(k) \leq \frac{C}{k^\alpha}$ and the second summand satisfies:

$$\begin{aligned}
\mathbb{E}_P\left[\left|\mathtt{C1}\left(X_{1:n}\right)\right|\right] &= \mathbb{E}_P\left[\sum_{k \geq \hat{K}_n+1} P^1(k)\ell(k)\right] \\
&\leq nC\mathbb{E}_P\left[\int_{\hat{K}_n}^{\infty} \frac{\ell\left(x\right)}{x^\alpha}\mathrm{d}x\right] \\
&= nC\mathbb{E}_P\left[\frac{1}{\hat{K}_n^{\alpha-1}}\int_1^\infty \frac{\ell\left(\hat{K}_n u\right)}{u^\alpha}\mathrm{d}u\right] \\
&\leq nC\mathbb{E}_P\left[\frac{1}{\hat{K}_n^{\alpha-1}}\right]\int_1^\infty \frac{\log\left(u\right)}{u^\alpha}\,\mathrm{d}u\left(1+o(1)\right) \\
&= O\left(n^{\frac{1}{\alpha}}\log n\right)
\end{aligned}$$

by Proposition (7) and Inequality (7).

By Theorem 2, every string $x \in \mathbb{N}_+^n$ satisfies

$$\left|\mathtt{C2}\left(x\right)\right| - n\sum_{k=1}^{\hat{K}_n} P^1(k)\log\frac{1}{P_1(k)} \leq \frac{\hat{K}_n}{2}\log n + 2.$$

Hence, the third summand is upper-bounded as:

$$\begin{aligned}
\mathbb{E}_P\left[\left|\mathtt{C2}\left(X_{1:n}\right)\right| - n\sum_{k=1}^{\hat{K}_n} P^1(k)\log\frac{1}{P_1(k)}\right] &\leq \frac{\mathbb{E}_P\left[\hat{K}_n\right]}{2}\log n + 2 \\
&= O\left(n^{\frac{1}{\alpha}}\log n\right)
\end{aligned}$$

which finishes to prove the theorem. ■

References

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998. ISSN 0018-9448.

S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Struct. & Algorithms*, 16:277–292, 2000.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, 2004. Ecole d'Ete de Probabilites de Saint-Flour XXXI.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. ISBN 0521841089.

B. Clarke and A. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Trans. Inform. Theory*, 36: 453–471, 1990.

B. Clarke and A. Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Stat. Planning and Inference*, 41:37–60, 1994.

T. Cover and J. Thomas. *Elements of information theory*. John Wiley & sons, 1991.

I. Csiszár. Class notes on information theory and statistics. University of Maryland, 1990.

I. Csiszár and P. Shields. Redundancy rates for renewal and other processes. *IEEE Trans. Inform. Theory*, 42(6): 2065–2072, 1996.

L. D. Davisson. Universal noiseless coding. *IEEE Trans. Information Theory*, IT-19:783–795, 1973. ISSN 0018-9448.

A. Dembo and O. Zeitouni. *Large deviation techniques and applications*. Springer, 1998.

L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer., 1996.

M. Drmota and W. Szpankowski. Precise minimax redundancy and regret. *IEEE Trans. Inform. Theory*, 50(11): 2686–2707, 2004. ISSN 0018-9448.

R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-00754-2. Revised reprint of the 1989 original.

P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory*, IT-21: 194–203, 1975. ISSN 0018-9448.

M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38(4):1258–1270, 1992. ISSN 0018-9448.

R. G. Gallager. *Information theory and reliable communication*. John Wiley & sons, 1968.

A. Garivier. A lower bound for the maximin redundancy in pattern coding. Technical report, Université Paris-Sud, 2006.

L. Gyorfi, I. Pali, and E. van der Meulen. On universal noiseless source coding for infinite source alphabets. *Eur. Trans. Telecommun. Relat. Technol.*, 4(2):125–132, 1993.

L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inform. Theory*, 40(1):267–271, 1994. ISSN 0018-9448.

D. Haussler. A general minimax result for relative entropy. *IEEE Trans. Inform. Theory*, 43(4):1276–1280, 1997. ISSN 0018-9448.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.*, 25(6):2451–2492, 1997. ISSN 0090-5364.

J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory*, 24(6):674–682, 1978. ISSN 0018-9448.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, 27(2): 199–207, 1981. ISSN 0018-9448.

A. Orlitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inform. Theory*, 50(10):2215–2230, 2004. ISSN 0018-9448.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inform. Theory*, 50(7):1469–1481, 2004. ISSN 0018-9448.

J. Rissanen and G. G. Langdon, Jr. Arithmetic coding. *IBM J. Res. Develop.*, 23(2):149–162, 1979. ISSN 0018-8646.

G. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory*, 52(5): 1939–1955, 2006. ISSN 0018-9448.

G. I. Shamir. A new redundancy bound for universal lossless compression of unknown alphabets. In *Proceedings of The 38th Annual Conference on Information Sciences and Systems-CISS*, pages 1175–1179, Princeton, New-Jersey, U.S.A., 2004.

P. Shields. Universal redundancy rates do not exist. *IEEE Trans. Inform. Theory*, 39:520–524, 1993.

Y. Shtarkov. Universal sequential coding of messages. *Probl. Inform. Transmission*, 23:3–17, 1987.

M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958. ISSN 0030-8730.

W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Probl. Inf. Transm.*, 34(2): 142–146, 1998.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math., Inst. Hautes Etud. Sci.*, 81:73–205, 1995.

E. Whittaker and G. Watson. *A course of modern analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1996. ISBN 0-521-58807-3. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition.

Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inform. Theory*, 43:646–656, 1997.

Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*, 46:431–445, 2000.

APPENDIX I
UPPER-BOUND ON MINIMAX REGRET

This sections contains the proof of the last inequality in Theorem 2.

The minimax regret is not larger than the maximum regret of the Krichevsky-Trofimov mixture over m-ary alphabet over strings of length n. The latter is classically upper-bounded by

$$\log\left(\frac{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})}{\Gamma(n+\frac{1}{2})\Gamma(\frac{m}{2})}\right) ,$$

as proved for example in (Csiszár, 1990).

Now the Stirling approximation to the Gamma function (See Whittaker and Watson, 1996, Chapter XII) asserts that for any $x > 0$, there exists $\beta \in [0, 1]$ such that

$$\Gamma(x) = x^{x-\frac{1}{2}}\mathrm{e}^{-x}\sqrt{2\pi}\mathrm{e}^{\frac{\beta}{12x}} .$$

Hence,

$$\log\left(\frac{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})}{\Gamma(n+\frac{1}{2})\Gamma(\frac{m}{2})}\right) = \left(n+\frac{m-1}{2}\right)\log\left(n+\frac{m}{2}\right) - n\log\left(n+\frac{1}{2}\right) - \frac{m-1}{2}\log\frac{m}{2} \tag{8}$$

$$-\left(n+\frac{m}{2}\right)+n+\frac{1}{2}+\frac{m}{2} \tag{9}$$

$$-\log\sqrt{2\pi}+\log\sqrt{2\pi}+\log\sqrt{2\pi}-\log\sqrt{\pi} \tag{10}$$

$$+\frac{\beta_1}{12\left(n+\frac{m}{2}\right)}-\frac{\beta_2}{12\left(n+\frac{1}{2}\right)}-\frac{\beta_3}{6m} \tag{11}$$

for some $\beta_1,\beta_2,\beta_3\in[0,1]$. Now, (9)+(10)+(11) is smaller than $\frac{1}{2}+\log\sqrt{2}+\frac{1}{12\left(n+\frac{m}{2}\right)}\leq 2$, and (8) equals:

$$\frac{m-1}{2}\log n+\left(n\log\frac{n+\frac{m}{2}}{n+\frac{1}{2}}-\frac{m-1}{2}\log e\right)+\left(\frac{m-1}{2}\log\frac{n+\frac{m}{2}}{\frac{m}{2}}+\frac{m-1}{2}\log e-\frac{m-1}{2}\log n\right)$$

But

$$n\log\frac{n+\frac{m}{2}}{n+\frac{1}{2}}=n\log\left(1+\frac{\frac{m-1}{2}}{n+\frac{1}{2}}\right)\leq n\frac{\frac{m-1}{2}}{n+\frac{1}{2}}\log e\leq\frac{m-1}{2}\log e,$$

and

$$\frac{m-1}{2}\log\frac{n+\frac{m}{2}}{\frac{m}{2}}+\frac{m-1}{2}\log e-\frac{m-1}{2}\log n=\frac{m-1}{2}\log\frac{\left(n+\frac{m}{2}\right)e}{\frac{nm}{2}}\leq 0$$

if $\left(n+\frac{m}{2}\right)e\leq\frac{nm}{2}$, that is $\frac{2}{m}+\frac{1}{n}\leq\frac{1}{e}$, which is satisfied as soon as m and n are both at least equal to 9. For the smaller values of $m,n\in\{2,\ldots,8\}$ the result can be checked directly.

APPENDIX II

LOWER BOUND ON REDUNDANCY FOR POWER-LAW ENVELOPES

In this appendix we derive a lower-bound for power-law envelopes using Theorem 5. Let α denote a real larger than 1. Let C be such that $C^{1/\alpha}>4$. As the envelope function is defined by $f(i)=1\wedge C/i^{\alpha}$, the constant $c(\infty)=\sum_{i\geq1}f(2i)$ satisfies

$$\frac{\alpha}{\alpha-1}\frac{C^{1/\alpha}}{2}-1\leq c(\infty)\leq\frac{C^{1/\alpha}}{2}+\frac{C}{(\alpha-1)2^{\alpha}}\left(\frac{C^{1/\alpha}}{2}\right)^{1-\alpha}.$$

The condition on C and α warrants that, for sufficiently large p, we have $c(p)>1$ (this is indeed true for $p>C^{1/\alpha}$).

We choose $p=an^{\frac{1}{\alpha}}$ for a small enough to have

$$\frac{(1-\lambda)C\epsilon}{(2a)^{\frac{1}{\alpha}}\,c(\infty)}>10,$$

so that condition $(1-\lambda)n\frac{f(2p)}{c(p)}>\frac{10}{\epsilon}$ is satisfied for n large enough. Then

$$R^{+}(\Lambda_f^n)\geq C(p,n,\lambda,\epsilon)\sum_{i=1}^{p}\left(\frac{1}{2}\log\frac{n\,(1-\lambda)\,\pi f(2i)}{2c(p)e}-\epsilon\right),$$

where $C(p,n,\lambda,\epsilon)=\frac{1}{1+\frac{(2a)^\alpha c(\infty)}{C\lambda^2}}\left(1-\frac{4}{\pi}\sqrt{\frac{5c(\infty)(2a)^\alpha}{(1-\lambda)C\epsilon}}\right)$, and

$$\begin{aligned}\sum_{i=1}^{p}\left(\frac{1}{2}\log\frac{n\,(1-\lambda)\,\pi f(2i)}{2c(p)e}-\epsilon\right) &\le \frac{p}{2}\log n-\frac{\alpha}{2}\sum_{i=1}^{p}\log i+\left(\frac{1}{2}\log\frac{(1-\lambda)\,\pi C}{2^{1+\alpha}c(\infty)e}-\epsilon\right)p\\
&=\frac{p}{2}\log n-\frac{\alpha}{2}\left(p\log p-p+o(p)\right)+\left(\frac{1}{2}\log\frac{(1-\lambda)\,\pi C}{2^{1+\alpha}c(\infty)e}-\epsilon\right)p\\
&=\frac{an^{\frac{1}{\alpha}}}{2}\log n-\frac{\alpha}{2}\left(an^{\frac{1}{\alpha}}\log a+\frac{a}{\alpha}n^{\frac{1}{\alpha}}\log n-an^{\frac{1}{\alpha}}+o\left(n^{\frac{1}{\alpha}}\right)\right)\\
&\qquad+\left(\frac{1}{2}\log\frac{(1-\lambda)\,\pi C}{2^{1+\alpha}c(\infty)e}-\epsilon\right)an^{\frac{1}{\alpha}}\\
&=\left(\frac{\alpha}{2}\left(1-\log a\right)+\frac{1}{2}\log\frac{(1-\lambda)\,\pi C}{2^{1+\alpha}c(\infty)e}-\epsilon+o(1)\right)an^{\frac{1}{\alpha}}.\end{aligned}$$

For a small enough, this gives the existence of a positive constant η such that $R^+(\Lambda_f^n)\ge\eta n^{\frac{1}{\alpha}}$.

APPENDIX III
PROOF OF PROPOSITION 7

Suppose that there exist k_0, c and C such that for all $k\ge k_0$, $\frac{c}{k^\alpha}\le p_k\le\frac{C}{k^\alpha}$.

For $0\le x\le\frac{1}{2}$, it holds that $-(2\log 2)x\le\log(1-x)\le -x$ and thus

$$\mathrm{e}^{-(2\log 2)nx}\le(1-x)^n\le\mathrm{e}^{-nx}.$$

Hence (as $p_k\le\frac{1}{2}$ for all $k\ge 2$) :

$$\begin{aligned}\sum_{k=k_0}^{\infty}\left(1-\left(1-\frac{c}{k^\alpha}\right)^n\right)\le\mathbb{E}[Z_n]&\le\sum_{k=1}^{\infty}\left(1-\left(1-\frac{C}{k^\alpha}\right)^n\right)\\
\sum_{k=k_0}^{\infty}\left(1-e^{-\frac{cn}{k^\alpha}}\right)\le\mathbb{E}[Z_n]&\le 1+\sum_{k=2}^{\infty}\left(1-e^{-\frac{(2\log 2)Cn}{k^\alpha}}\right)\\
\int_{k_0}^{\infty}\left(1-e^{-\frac{cn}{x^\alpha}}\right)\mathrm{d}x\le\mathbb{E}[Z_n]&\le 1+\int_{1}^{\infty}\left(1-e^{-\frac{(2\log 2)Cn}{x^\alpha}}\right)\mathrm{d}x.\end{aligned}$$

But, for any $t,K>0$, it holds that

$$\int_t^\infty\left(1-e^{-\frac{Kn}{x^\alpha}}\right)\mathrm{d}x=\frac{(Kn)^{1/\alpha}}{\alpha}\int_0^{\frac{Kn}{t^\alpha}}\frac{1-e^{-u}}{u^{1+1/\alpha}}\mathrm{d}u.$$

Thus, by noting that integral

$$A(\alpha)=\int_0^\infty\frac{1-e^{-u}}{u^{1+1/\alpha}}\mathrm{d}u,$$

is finite, we get

$$\frac{c^{1/\alpha}A(\alpha)}{\alpha}n^{1/\alpha}\left(1-o(1)\right)\le\mathbb{E}[Z_n]\le\frac{\left((2\log 2)C\right)^{1/\alpha}A(\alpha)}{\alpha}n^{1/\alpha}.$$

Appendix IV

Expected size of dictionary encoding

Assume that the probability mass function (p_k) satisfies $p_k \leq \frac{C}{k^\alpha}$ for $C > 0$ and all $k \geq 0$. Then, using Elias penultimate code for the first occurrence of each symbol in $X_{1:n}$, the expected length of the binary encoding of the dictionary can be upper-bounded in the following way. Let U_k be equal to 1 if symbol k occurs in $X_{1:n}$, and equal to 0 otherwise.

$$
\begin{aligned}
\mathbb{E}\left[|\Delta_n|\right] &= \mathbb{E}\left[\sum_{k=1}^{\infty} U_k \ell(k)\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left[U_k \ell(k)\right] \\
&\leq \sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{C}{k^\alpha}\right)^n\right) \ell(k) \\
&\leq 2\left(1 + \sum_{k=2}^{\infty} \left(1 - e^{-\frac{(2\log 2)Cn}{k^\alpha}}\right) \log k\right) \\
&\leq 2\left(1 + \int_1^{\infty} \left(1 - e^{-\frac{(2\log 2)Cn}{x^\alpha}}\right) \log x \, \mathrm{d}x\right) \\
&\leq 2\left(\frac{\left((2\log 2)\, Cn\right)^{1/\alpha}}{\alpha^2} \int_0^{(2\log 2)Cn} \frac{1 - e^{-u}}{u^{1+1/\alpha}} \log\left(\frac{(2\log 2)\, Cn}{u}\right) \mathrm{d}u\right) \\
&\leq T\, \frac{\left((2\log 2)\, Cn\right)^{1/\alpha}}{\alpha^2} \log n \int_0^{\infty} \frac{1 - e^{-u}}{u^{1+1/\alpha}} \mathrm{d}u
\end{aligned}
$$

for some positive constant T.

Acknowledgment

The authors wish to thank Lázslo Györfi for stimulating discussion and helpful comments.